UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

Institut Formation Doctorale

THÈSE DE DOCTORAT DE
L'UNIVERSITÉ PIERRE ET MARIE CURIE - PARIS 6

Spécialité

BIOMATHÉMATIQUES

Présentée par

MATTHIEU RESCHE-RIGON

Pour obtenir le grade de

DOCTEUR de L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Mesures d'influence individuelle pour modèles de régression en épidémiologie clinique

Soutenue le xx juin 2008

Devant le jury composé de :

Mme Sylvie Chevret, PU-PH	Directeur de thèse
Mr Jacques Bénichou, PU-PH	Rapporteur
Mr Jean-Christophe Thalabard, PU-PH	Rapporteur
Mr Alain-Jacques Valleron, PU-PH	Examinateur
Mr Alexandre Mebazaa, PU-PH	Examinateur

Université Pierre et Marie Curie - Paris 6
15, rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
E-mail : scolarite.doctorat@upmc.fr

Merci Sylvie pour ton aide, ta patience, ton soutien et surtout la confiance que tu m'as accordée tout au long de ce travail. Je te remercie encore vivement pour toutes les connaissances que tu me transmets et pour le plaisir que j'ai à travailler avec toi.

RÉSUMÉ

Mesures d'influence individuelle pour modèles de régression en épidémiologie clinique

Lors de l'analyse des données recueillies dans le cadre de projets de recherche clinique, les modèles statistiques intègrent habituellement de manière homogène l'information apportée par l'ensemble des observations. Les estimations obtenues peuvent cependant être modifiées par un nombre restreint voire par une seule observation, illustrant leur influence différente. Les mesures d'influence individuelle ont été proposées pour la quantifier, à la fois dans le but de détecter les influences et de mieux comprendre également les modèles auxquels elles s'appliquent. L'objet de ce travail est d'évaluer l'influence individuelle dans le cadre de modèles statistiques récents.

La première partie propose une mesure de l'influence individuelle locale pour le modèle de régression du risque instantané associé à la fonction d'incidence cumulée proposé par Fine et Gray pour l'analyse de données censurées en présence de compétition. La seconde partie du travail cherche à mettre en évidence l'influence individuelle des premiers sujets inclus dans un essai séquentiel de recherche de dose de phase I ou II utilisant la méthode de réévaluation séquentielle (MRS). Une adaptation de la MRS permettant de diminuer l'influence des premiers individus est enfin proposée.

MOTS CLÉS

Influence individuelle, Influence locale, Vraisemblance pondérée, Risques compétitifs, Schémas adaptatifs, Methode de Réévaluation Séquentielle (MRS)

ABSTRACT

Individual influence measures for regression models in clinical epidemiology

In analyzing data collected in clinical research, the statistical models usually incorporate information provided by all the observations. The estimates, however, can rely on a small (possibly one) number of observations, illustrating their different influence. Individual influence measures have been proposed that also allow an improved understanding of the models to which they apply. The purpose of this work was to evaluate the individual influence in the setting of recent statistical models.

The first section provides a measure of local influence in the proportional hazards model for the subdistribution function proposed by Fine and Gray to handle right censored data in the presence of competition. The second section of the work aims at highlighting the influence of the first individuals included in a dose finding trial (Phase I or II), using the Continual Reassessment Method (CRM). An adaptation of the CRM reducing the influence of first individuals is finally proposed.

KEY WORDS

Individual influence, Local influence, Weighted likelihood, Competing risks, Adaptive design, Continual Reassessement Method (CRM)

Ecole Doctorale Santé Publique et Sciences de l'Information Biomédicale

LABORATOIRE

Département de Biostatistique et Informatique Médicale, UMR 717 Inserm "Biostatistique et Epidémiologie Clinique", Université Denis Diderot–Paris 7, Hôpital Saint-Louis, 1, avenue Claude Vellefaux, 75010 Paris

Table des matières

Table des figures

Chapitre 1

Introduction

Les modèles mathématiques sont largement utilisés en sciences et notamment en biostatistique, le plus souvent dans un contexte de régression. Ils permettent au prix d'une description simplifiée, et donc partiellement fausse, de la réalité d'obtenir des informations sur celle-ci. Box a résumé cette contradiction par la célèbre formule : *All models are wrong, but some are useful* [1]. Si cette formule a le mérite d'être courte, elle n'indique pas le lien entre leur degré d'inexactitude et leur utilité, que souligne sa seconde version : *Remember that all models are wrong ; the practical question is how wrong do they have to be to not be useful* [1]. Tous les modèles sont donc faux, mais, en pratique, à partir de quel degré d'inexactitude perdent-ils leur utilité ? Répondre à cette question implique nécessairement une phase de remise en cause, de "critique" du modèle utilisé, notamment lors d'une analyse statistique. Box, Cook et Weisberg ont défini le "paradigme fondamental" de l'utilisation d'un modèle (*basic paradigm*) comme une succession d'allers et retours entre d'une part la formulation du problème et d'autre part les résultats de l'inférence [2]. Ainsi l'ajustement seul du modèle aux observations ne serait suffire. Une phase de critique à partir des résultats doit permettre de valider ou au contraire de remettre en cause modèle, hypothèses et postulats initiaux éventuels pour une nouvelle phase d'ajustement.

L'étude des perturbations du modèle prend toute son importance dans cette phase critique [3]. Le principe en est le suivant : si de légères variations dans la formulation du problème entraînent des modifications importantes des résultats, il y a lieu de s'interroger sur la validité du modèle dont ils sont issus [4]. C'est dans ce contexte que se place l'étude de l'influence des observations, à partir de mesures de la contribution de chaque individu aux résultats. Si ces mesures sont très hétérogènes, c'est à dire si les résultats ne dépendent que de très peu d'individus, le modèle est peut-être mal adapté aux données et peu à même de décrire correctement la réalité. Les mesures d'influence sont donc construites pour évaluer les effets sur l'inférence de modifications des données [5]. L'intérêt principal de ces mesures n'est pas de définir le meilleur modèle, mais de fournir une meilleure compréhension de la structure des données pour le modèle de régression considéré [6] et des informations quant à la qualité du lien entre modèle et conclusions [2].

Le concept d'influence individuelle a initialement été introduit par Cook en 1977 pour le modèle de régression linéaire [7]. Il consiste à évaluer les conséquences d'une perturbation des observations en modifiant le poids de chacune d'elles dans le modèle de régression. Pour chaque observation i, l'influence individuelle est mesurée par différence des estimations des paramètres du modèle lorsque tous les poids sont égaux à 1 sauf pour l'observation i (pour laquelle le poids est en règle nul) et celles obtenues lorsque tous les poids sont égaux à 1. Ceci revient donc à comparer les estimations en présence ou non d'une seule observation, d'où leur nom de méthodes "*delete-one*". Elles ont été rapidement étendues aux modèles de régression logistique [8] et au modèle de Cox [9]. Cependant, si ces techniques ont connu un succès rapide, elles pêchent par le coté radical de la perturbation ("en tout ou rien") étudiée et ont été suspectées de mal apprécier de faibles perturbations.

Pour mieux mesurer la sensibilité du modèle à de minimes variations du poids d'un individu, Cook a développé en 1986 le concept d'influence individuelle locale [3]. La mesure d'influence locale est alors définie par la direction de plus grande perturbation de l'estimateur du maximum de vraisemblance dans l'espace des poids des observations. Elle est maintenant considérée comme la mesure d'influence individuelle à privilégier sur l'approche *delete-one*, du fait de sa plus grande sensibilité aux perturbations du modèle [10].

Les mesures d'influence individuelle permettent donc d'identifier des observations ayant une influence sur les résultats [11, 48]. Mais leur intérêt principal, comme l'ont souligné Chen et Wang [6], est de fournir des éléments pour une meilleure compréhension des modèles. C'est dans cette perspective que s'est situé notre travail, dont l'objectif était d'étudier deux modèles statistiques de plus en plus fréquemment utilisés dans la modélisation statistique des données recueillies dans le cadre d'études d'épidémiologie clinique, par exemple en hématologie ou en réanimation. Il s'agit de la modélisation des risques compétitifs et des essais cliniques séquentiels de recherche de dose.

La notion de risque compétitifs est ancienne puisque son concept a été introduit par Bernoulli au XVIII^{me} siècle. Elle définit une situation où un individu est exposé simultanément au risque de survenue de plusieurs événements, la survenue de l'un annulant ou modifiant la probabilité de survenue des autres. Dans l'analyse des données de protocole de recherche clinique, de telles situations sont fréquentes, par exemple en greffe de moelle. Ainsi, un patient allogreffé peut développer une maladie du greffon contre l'hôte (*GvHD*), mais la survenue d'un décès précoce, avant *GvHD*, en empêche toute survenue : on dit que ces deux risques sont "en compétition". Des modèles de régression adaptés à ce contexte ont été proposés, tels le modèle cause-spécifique de Cox [11] et plus récemment le modèle de Fine et Gray pour la fonction de risque de sous-répartition [12]. L'objectif de la première partie de ce travail de recherche a été de développer une mesure d'influence individuelle locale pour ce dernier modèle, en mettant en évidence les différences qu'il présente avec le premier dans la modélisation des risques compétitifs.

La deuxième partie de ce travail s'est intéressée à l'influence individuelle dans la modélisation des données recueillies dans le cadre d'essais séquentiels adaptatifs, en prenant comme exemple la Méthode de Réévaluation Séquentielle (MRS) ou *Continual Reassessment Method, CRM*. Cette méthode a été proposée en 1990 pour conduire et analyser séquentiellement des essais de phase I en cancérologie [13]. Elle se caractérise par une modélisation mathématique paramétrique de la relation dose toxicité, dont les paramètres sont re-estimés séquentiellement au fur et à mesure de l'acquisition des observations. Ces estimations séquentielles interviennent dans la règle d'attribution des doses pour les malades suivants. Dans ce contexte de schéma adaptatif, il est impossible de mesurer l'influence individuelle via les approches *delete-one* ou les mesures d'influence locale. En effet, le retrait d'une observation conduit à modifier potentiellement la conduite ultérieure de l'essai. D'autres approches doivent donc être développées s'inspirant des techniques d'analyses de sensibilité. La seconde partie de ce travail de recherche a eu pour objectif de mettre en évidence l'influence relative des premières observations dans le cadre de la MRS, en utilisant une inférence basée sur la vraisemblance pondérée.

Dans un premier temps, nous rappellerons le principe des mesures d'influence individuelle pour les modèles de régression linéaire. Les développements propres à ce travail de recherche sont présentés dans les chapitres suivants, avec pour chacun, une brève introduction au problème soulevé, un exemple illustratif, la présentation des méthodes proposées, puis l'article associé. Enfin, le dernier chapitre présente quelques éléments de discussion.

Pour mieux faire référence à la littérature anglo-saxonne, très abondante dans ce contexte, et par souci de clarté, nous avons délibérément choisi de faire figurer entre parenthèses et en italique, les termes anglais correspondants.

Chapitre 2

Mesure d'influence individuelle pour les modèles de régression linéaire

2.1 Rappels

Soit Y une variable réponse et $X = (X_1, X_2, \ldots, X_p)$ un p-vecteur de covariables explicatives. Posons $X^T\beta = \beta_1 X_1 + ... + \beta_p X_p$, le prédicteur linéaire de X, où $\beta = (\beta_1, \beta_2, ..., \beta_p)$ est un p-vecteur de paramètres dits coefficients de régression. Le modèle linéaire permet de relier au vecteur de covariables X l'espérance conditionnelle de Y sachant la réalisation de X [14, 15]. Il s'écrit $E(Y|X) = \alpha + X^T\beta$, où X^T est la transposée de X et $\alpha \in \mathbb{R}$ l'intercept. Par souci de simplification, on se placera par la suite dans le cas où $\alpha = 0$

La régression linéaire cherche à partir d'un échantillon de n observations indépendantes de (Y, X), entachées d'erreurs de mesure, à définir le modèle mathématique linéaire (supposé les décrire) qui ajuste aux mieux les observations. On notera $x = \{(x_{1i}, \ldots, x_{pi}), i = 1, \ldots, n\}$, la matrice $n \times p$ des covariables observées, considérées comme fixes, et $y = (y_1, y_2, ...y_n)$ le vecteur des réponses observées considérées comme les réalisations d'un vecteur aléatoire Y. Le modèle s'écrit alors $y = x^T\beta + \varepsilon$, où $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ est un n-vecteur d'erreurs inobservables. Ces erreurs sont supposées indépendantes et distribuées selon une loi $\mathcal{N}$ormale de moyenne nulle et de variance σ^2. L'estimation des coefficients de régression β repose sur la méthode des moindres carrés, qui consiste à minimiser la somme des carrés des résidus du modèle, c'est à dire la distance entre les observations et le modèle. Le résidu ordinaire e_i de l'observation i est défini par la différence entre la réponse observée de l'individu i, y_i, et celle prédite par le modèle $\hat{y}_i = x_i^T\hat{\beta}$, où $\hat{\beta} = (x^T x)^{-1}x^T y$ est l'estimateur des moindres carrés de β. Il est équivalent ici à l'estimateur du maximum de vraisemblance. L'étude graphique des résidus, parfois studentisés (*standardized or studentized residuals*), contre le rang des observations ou les valeurs de x permet de vérifier l'hypothèse de normalité des erreurs ε.

2.2 Mesures d'influence individuelle

Soit n, le nombre d'observations et p la dimension du vecteur de coefficients de régression linéaire à estimer.

Pour étudier l'influence individuelle dans le modèle linéaire, les résidus ont tout d'abord été utilisés. Malheureusement, ils reflètent mal l'influence individuelle sur les résultats, comme l'illustre la figure 2.1. Dans cet exemple précédemment décrit par Cook et Weisberg [2], la droite de régression estimée à partir de l'ensemble des observations diffère de façon importante de celle estimée en excluant une seule observation (figure 2.1.a). Cependant, l'étude graphique des résidus (figure 2.1.b) ne permet pas de détecter cette observation.

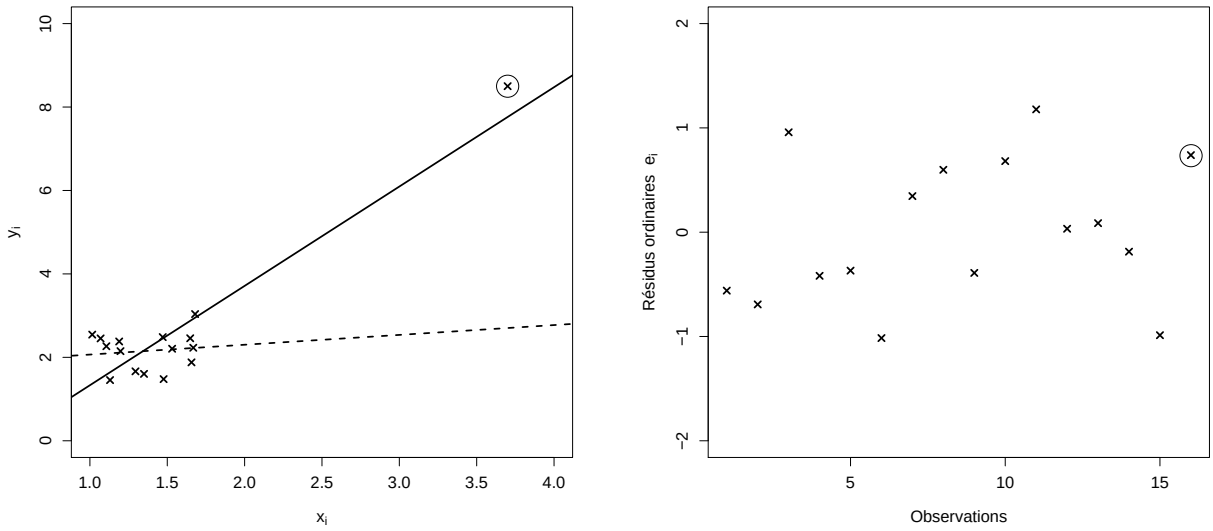


Fig 2.1.a Fig 2.1.b

FIG. 2.1 – **Illustration sur un exemple des limites des résidus dans l'identification de données influençantes dans une régression linéaire**. Fig 2.1.a : Estimation de la droite de régression linéaire avec l'ensemble des 16 observations (droite en trait plein) ou en excluant l'observation cerclée (droite en pointillé). Fig 2.1.b : Résidus ordinaires dans le modèle linéaire des 16 observations en fonction du rang de l'observation

D'autres mesures d'influence individuelles sont donc apparues nécessaires. On distingue les mesures p-dimensionnelles comme par exemple les *Diagnostics* basés sur l'approche *delete-One*, les mesures d'influence globale comme la distance de Cook, et les mesures d'influence locale comme le déplacement de la vraisemblance.

2.2.1 *Diagnostics* basés sur l'approche *delete-One*

Les premières mesures d'influence dites *case deletion diagnostics* fournissent une mesure de l'influence de chaque individu sur chacun des p coefficients de régression [16, 8]. Leur principe est simple. Il consiste à exclure un individu de l'échantillon, ré-estimer les p coefficients de régression du modèle et comparer les estimations obtenues à celles réalisées avec l'ensemble des individus (*delete-one approach*).

Si on note $\hat{\beta_{(i)}}$ l'estimation des coefficients du modèle en excluant l'observation i (on gardera par la suite cette notation pour toutes les estimations faites sans l'observation i), on obtient un p-vecteur mesurant les variations de l'estimation des coefficients pour chaque individu i. Cette mesure est connue sous le nom de $dfbeta_i = \hat{\beta} - \hat{\beta_{(i)}}$ [16, 17]. On définit aussi $dfbetas_i = \frac{\hat{\beta} - \hat{\beta_{(i)}}}{SE(\hat{\beta_{(i)}})}$ qui correspond à la mesure précédente standardisée. Tracer le graphe des $dfbetas_i$ contre le rang des observations i ou contre les valeurs x_i permet de détecter graphiquement les observations influençantes [10, 15]. Cependant, si ces mesures apportent de l'information sur l'impact d'un individu sur les estimations, elles ont l'inconvénient de fournir une mesure multidimensionnelle de l'influence individuelle.

On définit également les $dffit_i = \hat{y} - \hat{y_{(i)}}$ qui correspondent aux différences des prédictions effectuées avec ou sans le sujet i [16]. On définit de façon similaire $dffits_i$ en standardisant les $dffit_i$. Là encore, il est recommandé une détection graphique des individus influençants. Néanmoins, l'influence est souvent mesurée par $|dffits_i| > 2\sqrt{\frac{p+1}{n-p-1}}$ [16, 15]. Ces mesures présentent l'avantage de résumer l'influence de l'individu i par un scalaire, malheureusement la standardisation dépendant de l'individu i (comme pour les $dfbetas_i$), les comparaisons entre individus sont difficiles.

2.2.2 Distance de Cook

Des mesures d'influence globale, c'est à dire résumant l'influence des individus par une mesure unidimensionnelle, ont alors été développées. L'objectif était double : (i) normer l'espace à p dimensions des $dfbeta$, et (ii) obtenir une norme qui ne dépende pas de l'observation i. La plus utilisée est l'une de celles décrites par Cook, connue sous le nom de distance de Cook [2, 3]. Classiquement notée D_i, elle se définit par :

$$D_i = \frac{(\hat{\beta} - \hat{\beta_{(i)}})^T (x^T x)(\hat{\beta} - \hat{\beta_{(i)}})}{p\hat{\sigma}^2} = \frac{(\hat{y} - \hat{y_{(i)}})^T (\hat{y} - \hat{y_{(i)}})}{p\hat{\sigma}^2}. \tag{2.1}$$

Elle présente l'avantage d'être invariante aux changements d'échelle, de dépendre de $\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-p}$ et d'être directement reliée à la notion de déplacement de vraisemblance.

Soit $w = (w_1, ..., w_n)$ un n-vecteur de poids et $L(\beta, w)$, la fonction de log-vraisemblance pondérée :

$$L(\beta, w) = \sum_{i=1}^{n} w_i \log(f(y_i, x_i, \beta)) \tag{2.2}$$

où $f(y_i, x_i, \beta)$ est la densité des observations. On définit $\hat{\beta(w)}$ comme l'estimation de β obtenue en maximisant cette vraisemblance pondérée. Définissons le déplacement de vraisemblance $LD(w)$ par :

$$LD(w) = 2[L(\hat{\beta}) - L(\hat{\beta(w)})] \tag{2.3}$$

A noter que $\hat{\beta_{(i)}}$ est alors un cas particulier de $\hat{\beta(w)}$ pour $w_j = 1$ pour $j \neq i$ et $w_i = 0$.
On définit :

$$D(w) = \frac{(\hat{\beta} - \hat{\beta(w)})^T (x^T x)(\hat{\beta} - \hat{\beta(w)})}{p\hat{\sigma}^2} = \frac{(\hat{y} - \hat{y(w)})^T (\hat{y} - \hat{y(w)})}{p\hat{\sigma}^2} \tag{2.4}$$

Pour le modèle linéaire, on peut alors écrire :

$$pD(w) = \frac{(y - \hat{y(w)})^T (y - \hat{y(w)}) - (y - \hat{y})^T (y - \hat{y})}{\hat{\sigma}^2} = 2[L(\hat{\beta}) - L(\hat{\beta(w)})] \tag{2.5}$$

et donc,

$$D_i = \frac{2}{p}[L(\hat{\beta}) - L(\hat{\beta_{(i)}})] \tag{2.6}$$

Cette écriture de la distance de Cook (équation 2.6) permet de la définir pour l'ensemble des modèles linéaires généralisés [18, 2]. Il est en effet possible par une approximation d'ordre 2 du déplacement de vraisemblance d'écrire :

$$D_i \simeq \frac{1}{p}(\hat{\beta} - \hat{\beta_{(i)}})^T \frac{\partial^2 L(\beta)}{\partial \beta^2} (\hat{\beta} - \hat{\beta_{(i)}}) = \frac{1}{p}(\hat{\beta} - \hat{\beta_{(i)}})^T I (\hat{\beta} - \hat{\beta_{(i)}}) \tag{2.7}$$

où I est la matrice d'information de Fisher en $\hat{\beta}$. Le calcul de la distance de Cook devient alors simple et ce d'autant plus que des calculs approchés de $\hat{\beta} - \hat{\beta_{(i)}}$ basés sur des approximations d'ordre 1 ont été proposés [18].

En 1995, Lawrance proposa une mesure dérivée de la distance de Cook adaptée à la mesure de l'influence d'une paire d'observations [19]. Il définit alors la notion d'influence jointe, pointant le fait que l'influence de certains sujets peut être grandement modifiée par la présence ou non d'un autre sujet.

Si la distance de Cook a rapidement été adoptée et recommandée, il apparait que cette mesure manque de sensibilité dans l'analyse de perturbations plus fines du modèle, comme Cook l'a précédemment décrit (figure 2.2) [3]. Du point de vue de la distance de Cook, les situations A et B décrites sur la figure sont équivalentes pour l'influence du sujet i. Il semble pourtant que la situation A soit plus stable autour de $w_i = 1$ que la situation B. De cette observation, Cook déduisit la nécessité de construire une mesure d'influence locale, c'est à dire au voisinage proche de $w_0 = 1$ le vecteur unité de dimension n.

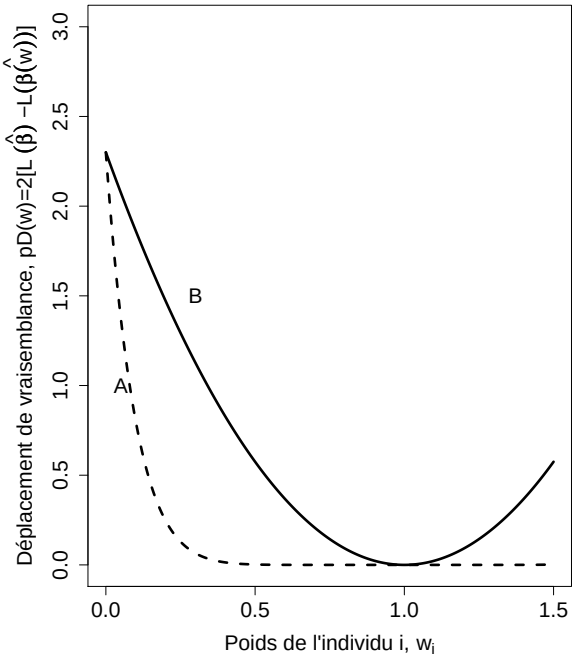


FIG. 2.2 – **Deux exemples A et B de modification du déplacement de vraisemblance en fonction du poids w_i accordée à une observation i.** $pD_A(1) = pD_B(1) = 0$ et $pD_A(0) = pD_B(0)$. L'analyse apparait plus sensible dans le cas B, $\forall\ w,\ pD_B(w) - pD_A(w) \geq 0$.

2.2.3 Influence locale

Le déplacement de vraisemblance autour de w_0 s'écrit avec $\hat{\beta} = \beta(\hat{w}_0)$:

$$\begin{aligned} LD(w) &= 2[L(\hat{\beta}) - L(\beta(\hat{w}))] \\ &\simeq (\hat{\beta} - \beta(\hat{w}))^T I(\hat{\beta} - \beta(\hat{w})) \end{aligned} \tag{2.8}$$

En posant $\hat{\beta} - \beta(\hat{w}) = \left[\frac{\partial\beta(\hat{w})}{\partial w}\right]_{\hat{\beta},w_0}(w - w_0)$, puis $\frac{\partial\beta(\hat{w})}{\partial w} = I^{-1}\frac{\partial U(\beta,w)}{\partial w}$, avec U le vecteur score, on obtient :

$$LD(w) \simeq (w - w_0)^T \Delta^T I^{-1} \Delta (w - w_0) \tag{2.9}$$

où Δ est la matrice $p \times n$ définie par : $\Delta = \left[\frac{\partial U(\beta,w)}{\partial w}\right]_{\hat{\beta},w_0}$.

Considérant un espace à $n + 1$ dimensions, $(w, LD(w))$ représente une surface dénommée graphique d'influence (*influence graphic*) [3, 20, 21]. Le principe proposé par Cook, en 1986 est de s'intéresser à la courbure de cette surface au voisinage de w_0. Soit C_l, la courbure de la surface dans une direction donnée l, on a $C_l = 2|l^T \frac{\partial^2 LD(w)}{\partial w \partial w} l|$, ou encore $C_l = 2|l^T \Delta^T I^{-1} \Delta l|$ [3, 22].

Cook s'intéressant à la direction qui maximise le déplacement de vraisemblance, c'est à dire au vecteur propre de norme 1, l_{max}, de la plus grande valeur propre $|\gamma_{max}|$ de la matrice $\Delta^T I^{-1}\Delta$, il définit $|l_{max}|$ comme la mesure de l'influence locale [3, 20]. L'influence de chaque observation i correspond à la valeur absolue de la coordonnée $(l_{max})_i$ de l_{max}. Il est alors recommandé de tracer l'influence locale de i, $(l_{max})_i$, contre i, la détection des individus les plus influençants se faisant là aussi graphiquement. Cook propose par ailleurs une valeur seuil de 1 pour $|\gamma_{max}|$, c'est à dire une courbure C_{max} supérieure à 2, comme règle de détection d'une sensibilité locale notable du modèle [3, 23].

D'autres auteurs pointeront l'importance des autres vecteurs propres de la matrice $\Delta^T I^{-1}\Delta$. Ainsi, Wu et Luo proposeront, en 1993, de tracer (en plus de $|l_{max}|$) la valeur absolue du vecteur propre correspondant à la deuxième plus grande valeur propre en valeur absolue [24]. Enfin, Poon et Poon ont proposé de considérer tous les n vecteurs propres l_j $(j = 1, ..., n)$ et leur valeur propre associée γ_j [22, 21]. Ils définissent alors pour chaque observation i une contribution aggrégée à chaque vecteur propre (*aggregate contribution*), notée $C_{e_i} = \sum_{j=1}^{r} \gamma_j \times (l_j)_i^2$, où r est le nombre de valeurs propres non nulles. Par ailleurs, ils conseillent de s'intéresser à tous les vecteurs propres associés à des valeurs propres supérieures à $1/r$. De ces travaux, Zhu at Zhang ont développé, en 2004, une statistique (fonction des C_{e_i}) permettant de tester la validité du nombre de paramètres du modèle [25].

La notion d'influence locale de Cook a par la suite été étendue à l'ensemble des modèles linéaires généralisés [20]. Des développement plus récents ont été proposés dans le cadre des modèles linéaires à effets mixtes [26, 27], des modèles pour données longitudinales [28] ou les modèles de calibration [29].

Chapitre 3

Influence et modèles de survie en présence de compétition

3.1 Modèles de survie en présence de compétition

3.1.1 Données de survie

Dans de nombreuses situations d'épidémiologie clinique, la survie, c'est à dire le délai de survenue du décès, est le critère d'intérêt mesurant le devenir des malades. Son analyse nécessite de recueillir non seulement l'information concernant la survenue ou non du décès, mais surtout le délai de survenue T du décès, calculé depuis une date d'origine (diagnostic ou début du traitement, par exemple) [10, 30].

Le terme "données de survie", devenu générique, s'est secondairement étendu à tout délai de survenue d'un événement en "tout ou rien", tel par exemple une rechute, un sevrage médicamenteux, ou une sortie de l'hôpital. A noter que si les anglo-saxons utilisent le terme de *survival data* pour décrire ces données, ils emploient également des termes plus généraux tels *time to event data* ou *time to failure data*.

3.1.1.1 Loi de probabilité

Une donnée (un délai) de survie est décrit(e) par une variable aléatoire T positive, dont la particularité est d'avoir en règle une distribution asymétrique à droite. Soit $f(t)$ sa densité.

La loi de probabilité de T est décrite préférentiellement par une fonction de survie $S(t)$ (*survival fonction*), qui est définie en t comme la fraction d'individus encore en vie en t. C'est donc le complémentaire à 1 de la fonction de répartition de T, $F(t)$:

$$S(t) = P(T > t) = 1 - F(t) \tag{3.1}$$

Une autre fonction particulièrement utilisée pour décrire la loi de probabilité de T est la fonction de risque instantané $h(t)$ (*hazard fonction*), définie comme la densité de probabilité de T conditionnellement à la survie en t :

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\} \tag{3.2}$$

Ces trois fonctions, $S(t)$, $F(t)$ et $h(t)$, sont mathématiquement liées. On a ainsi :

$$h(t) = \frac{f(t)}{S(t)} = \frac{1}{1 - F(t)} \frac{\mathrm{d}F(t)}{\mathrm{d}t} = -\frac{\mathrm{d}(\log(S(t)))}{\mathrm{d}t} \tag{3.3}$$

ou, de façon équivalente :

$$S(t) = \exp\left(- \int_0^t h(u) \mathrm{d}u \right) \tag{3.4}$$

3.1.1.2 Censure à droite

La caractéristique principale des données de survie, outre l'asymétrie à droite de leur distribution, est d'être potentiellement censurées. Si les données peuvent être censurées de plusieurs façons, on s'intéressera uniquement dans ce travail à la censure à droite. Un délai de survie est dit censuré à droite si l'événement d'intérêt n'a pas été observé au terme de l'observation : le délai de survie de l'individu est donc strictement supérieur au délai d'observation.

Si les différents mécanismes générant des censures à droite permettent de définir des censures fixes (type I), séquentielles (type II) et aléatoires (type III), le mécanisme de censure à doite générant les données recueillies dans le cadre d'études cliniques est en règle aléatoire. On définit donc, pour chaque individu i, $i = 1, ..., n$, deux variables aléatoires, le délai d'événement T_i, et le délai de censure C_i. Les observations se résument à $\{(T_i^*, \delta_i), i, \ldots, n\}$, où $T_i^* = T_i \wedge C_i$ et $\delta_i = \mathbf{1}_{[T_i \leq C_i]}$, $\mathbf{1}_{[\ ]}$ étant la fonction indicatrice.

L'inférence statistique sur la loi de T en présence de censure est basée sur une fonction de vraisemblance partielle, ignorant l'information sur la loi de T contenue dans C : la censure doit être "non informative" sur la loi de T. En termes probabilistes, le délai de censure C est supposé indépendant du délai de survie T. En épidémiologie clinique, pour s'assurer de l'indépendance entre délai de censure et délai de survie, on choisit de fixer une "date de point" (*reference date*), date à laquelle le statut de chaque sujet est évalué : tout événement survenu postérieurement à la date de point est censuré à cette date. Cette censure, dite "administrative", assure ainsi la validité de l'inférence statistique. Si, par contre, le statut d'un individu est inconnu à la date de point car son suivi s'est interrompu avant cette date, on dit que le sujet est "perdu de vue" (*withdrawal*). Dans ce cas, l'hypothèse d'indépendance entre la censure et l'événement d'intérêt peut être mise en doute, surtout si les causes d'interruption de suivi ne sont pas connues et/ou qu'elles sont nombreuses (Figure 3.1). On se placera par la suite sous l'hypothèse de censure non informative.

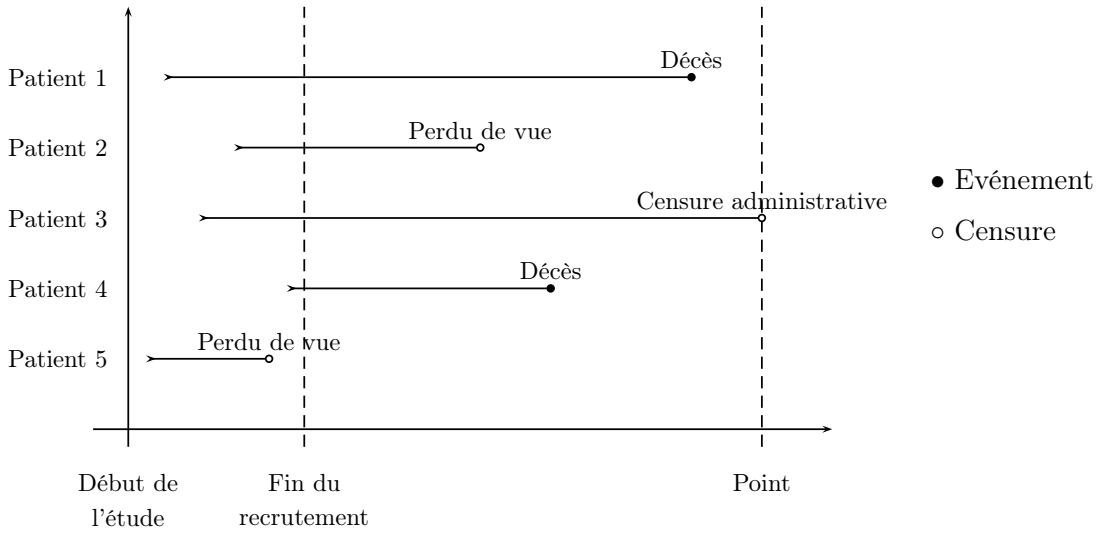


FIG. 3.1 – Représentation schématique de 5 données de survie

3.1.1.3 Modèle de régression semi-paramétrique de Cox

Cox proposa en 1972 un modèle de régression semi-paramétrique pour modéliser la fonction de risque instantané comme une fonction multiplicative d'un p-vecteur de covariables $X = (X_1, X_2, ..., X_p)$. Il s'écrit :

$$h(t, X) = h_0(t) \times \exp(X^T \beta) \tag{3.5}$$

où $\beta = (\beta_1, \beta_2, ..., \beta_p)$ est un p-vecteur de coefficients de régression à estimer [11] et $h_0(t)$ est une fonction de risque instantané de base non spécifiée, qui ne dépend ni de X ni de β.

Ce modèle semi-paramétrique implique, en présence de covariables X fixes, indépendantes du temps, des risques proportionnels au cours du temps (*proportional hazards*).

L'estimation des coefficients de régression s'obtient, sous l'hypothèse de censure non-informative, en maximisant la fonction de log-vraisemblance partielle (*log partial likelihood*) qui s'écrit :

$$L(\beta) = \sum_{i=1}^{n} \delta_i \left(X_i^T \beta - \log \left(\sum_{j \in R_i} \exp \left(X_j^T \beta \right) \right) \right) \tag{3.6}$$

où $R_i = \{l : T_l^* \geq T_i\}$ (*Risk set*), représente l'ensemble des individus à risque de présenter l'événement au temps T_i.

3.1.2 Données de survie en présence de compétition

Dans certains cas, le patient peut être exposé simultanément à la survenue de plusieurs événements, la survenue de l'un annulant ou modifiant la probabilité de survenue des autres [31, 32, 33]. Cette situation où un sujet est exposé simultanément à plusieurs risques exclusifs définit une situation dite de risques compétitifs ou en compétition (*Competing risks*) (Figure 3.2). C'est par exemple le cas en cancérologie où la rechute avant décès peut entrer en compétition avec le décès avant rechute, mais aussi en réanimation où la sortie vivant de réanimaton peut être considérée comme un risque en compétition du décès en réanimation.

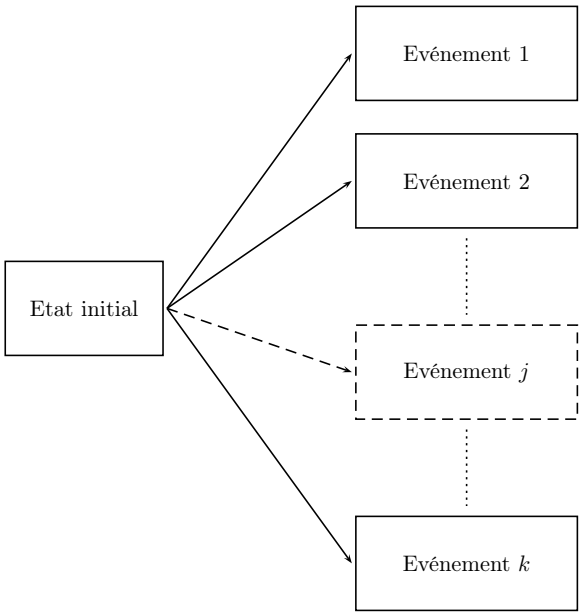


FIG. 3.2 – Représentation schématique d'une situation de risques compétitifs

On considèrera par la suite une situation où chaque individu est exposé au risque de survenue de k types d'événements (ou d'échecs) différents, mutuellement exclusifs : chaque individu exposé à ces k risques en compétition ne peut en présenter qu'un seul.

3.1.2.1 Notations

Soit T_i, le délai de survenue de l'événement (quelqu'il soit), et $\varepsilon_i \in \{1, ..., k\}$ le type d'événement observé en T_i. Soit C_i, le délai de censure. Les observations sont $\{(T_i^*, \varepsilon_i^*), i, \ldots, n\}$, où $T_i^* = T_i \wedge C_i$ et $\varepsilon_i^* = \mathbf{1}_{[T_i \leq C_i]} \times \varepsilon_i$.

3.1.2.2 Loi de probabilité

Soit $j \in \{j = 1, ..., k\}$, l'événement d'intérêt. Diverses fonctions pour décrire la loi de probabilité de T_j, délai de survenue de l'événement de type j, ont été décrites :

- La fonction de sous-répartition (*subdistribution fonction*) ou incidence cumulée (*cumulative incidence fonction, CIF*) de l'événement de type j, définie par

$$F_j(t) = P(T < t, \varepsilon = j) \tag{3.7}$$

- La fonction de risque instantané spécifique de l'événement de type j (*cause-specific hazard*), définie par le risque instantané d'événement j conditionnellement au fait de n'avoir fait aucun événement précédemment :

$$h_j(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t, \varepsilon = j | T \geq t)}{\delta t} \right\} \tag{3.8}$$

Cependant, contrairement à la situation précédente des données de survie où l'individu n'est exposé qu'à un seul type d'événement, il n'existe pas de relation simple entre ces deux fonctions. En effet, la fonction $h_j(t)$ dépend non seulement de la fonction de sous-répartition de l'événement j, $F_j(t)$, mais aussi de toutes les autres fonctions de sous-répartition [34] :

$$h_j(t) = \frac{1}{1 - \sum_{s=1}^{k} F_s(t)} \frac{\mathrm{d}F_j(t)}{\mathrm{d}t} \tag{3.9}$$

3.1.3 Modèles de régression en présence de compétition

Deux approches de régression ont été proposées, l'une privilégiant la fonction de risque cause-spécifique, l'autre la fonction de sous-répartition. L'approche privilégiant le risque cause-spécifique revient à évaluer l'effet d'une covariable sur un événement en supposant que les autres types d'événement ont été supprimés. L'approche privilégiant la fonction de sous-répartition considère un univers où tous ces événements sont possibles. De ces deux approches découlent deux modèles de régression semi-paramétriques différents, mais cependant tous deux fortement apparentés au modèle de Cox.

Par commodité de notation, nous considérerons dans le reste de ce travail que l'événement d'intérêt est l'événement de type 1.

3.1.3.1 Modèle pour la fonction de risque cause-spécifique

Le modèle semi-paramétrique de Cox pour la fonction de risque spécifique de l'événement de type 1 s'écrit :

$$h_1(t, X) = h_{10}(t) \exp(X^T \beta^C) \tag{3.10}$$

où $X = (X_1, X_2, ..., X_p)$ est un p-vecteur de covariables, $\beta^C = (\beta_1^C, \beta_2^C, ..., \beta_p^C)$ un p-vecteur de coefficients de régression à estimer, et $h_{10}(t)$ une fonction de risque instantané de base non spécifiée spécifique de l'événement 1, qui ne dépend ni de X, ni de β^C.

La fonction de log-vraisemblance partielle associée au modèle s'écrit alors :

$$L(\beta^C) = \sum_{i=1}^{n} \mathbf{1}_{[\varepsilon_i=1]} \left(X_i^T \beta^C - \log \left(\sum_{j \in R_i^C} \exp\left(X_j^T \beta^C\right) \right) \right) \tag{3.11}$$

où $\mathbf{1}_{[\,]}$ représente l'indicatrice et $R_i^C = \{j : (T_j^* \geq T_i)\}$ l'ensemble des individus à risque d'expérimenter l'événement de type 1 au temps T_i, c'est à dire l'ensemble des individus n'ayant expérimenté aucun des k événements possibles au temps T_i.

A noter que le lien entre $h_1(t)$ et la fonction de sous-répartition $F_1(t)$ n'étant pas direct, l'effet d'une covariable sur $h_1(t)$ peut différer de celui sur $F_1(t)$.

3.1.3.2 Modèle pour la fonction de risque de sous-répartition

Fine et Gray [12] ont proposé un modèle à risques proportionnels pour la fonction de risque de sous-répartition, $\lambda_1(t)$, fonction de risque instantané associée à la fonction de sous-répartition spécifique de l'événement de type 1 (*subdistribution hazard function*), décrite par Gray [31]. Elle s'écrit :

$$\lambda_1(t) = -\frac{\mathrm{d}\log(1 - F_1(t))}{\mathrm{d}t} \tag{3.12}$$

$$= \lim_{\delta t \to 0} \frac{1}{\delta t} P\{t \leq T \leq t + \delta t, \varepsilon = 1 | T \geq t \cup (T \leq t \cap \varepsilon \neq 1)\} \tag{3.13}$$

La fonction $\lambda_1(t)$ apparaît donc comme la fonction de risque instantané d'une variable aléatoire impropre $T^\bullet = \mathbf{1}_{[\varepsilon=1]} \times T + \{1 - \mathbf{1}_{[\varepsilon=1]}\} \times \infty$.

A l'instar de Cox, Fine et Gray ont proposé un modèle semi paramétrique à risques proportionnels pour décrire cette fonction de risque instantané selon un p-vecteur de covariables $X = (X_1, X_2, ..., X_p)$. Le modèle s'écrit :

$$\lambda_1(t, X) = \lambda_{10}(t) \exp\left(X^T \beta^F\right) \tag{3.14}$$

où $\beta^F = (\beta_1^F, \beta_2^F, ..., \beta_p^F)$ est un p-vecteur de coefficients de régression à estimer et $\lambda_{10}(t)$ est la fonction de risque instantané de base associée à la fonction de sous-répartition spécifique de l'événement 1, qui ne dépend ni de X, ni de β^F.

En l'absence de censure, l'estimation des coefficients β s'effectue par maximisation de la fonction de log-vraisemblance partielle :

$$L(\beta^F) = \sum_{i=1}^{n} \mathbf{1}_{[\varepsilon_i=1]} \left(X_i^T \beta^F - \log \left(\sum_{j \in R_i^F} \exp\left(X_j^T \beta^F\right) \right) \right) \tag{3.15}$$

où $R_i^F = \{j : (T_j \geq T_i) \cup (T_j \leq T_i \cap \varepsilon_j \neq 1)\}$ représente l'ensemble des individus à risque d'événement 1 au temps T_i, c'est-à-dire l'ensemble des individus n'ayant pas développé d'événement quelqu'il soit avant T_i ou ayant expérimenté un autre événement que celui d'intérêt avant T_i.

En présence de censure à droite, la vraisemblance prend en compte la distribution des délais de censure selon une méthode de pondération inverse. Sous l'hypothèse d'une censure non informative, les observations des individus ayant expérimenté un autre événement que celui d'intérêt sont alors pondérées par leur probabilité estimée $v_{..}$ d'être censurées. La fonction de log-vraisemblance devient :

$$L(\beta^F, v_{ji}) = \sum_{i=1}^{n} \mathbf{1}_{[\varepsilon_i=1]} \left(X_i^T \beta^F - \log \left(\sum_{j \in R_i^F} v_{ji} \exp\left(X_j^T \beta^F\right) \right) \right) \tag{3.16}$$

où $R_i^F = \{j : (T_j \geq T_i) \cup (T_j \leq T_i \cap \varepsilon_j \notin \{0,1\})\}$, et $v_{ji} = \hat{G}(T_i)/\hat{G}(T_j \wedge T_i)$, et $\hat{G}$ étant l'estimation par la méthode de Kaplan-Meier de la fonction de répartition des délais de censure, $G(t) = Pr(C > t)$ [12]. En fait, $v_{ji} = 1$ pour tous les indvidus vérifiant $T_j \geq T_i$ et $v_{ji} < 1$ pour les individus qui ont expérimenté un autre événement que celui d'intérêt et qui restent donc à risque. En l'absence de censure, $\forall\ (i,j),\ v_{ji} = 1$, on retrouve l'expression de la log vraisemblance (3.15).

3.2 Illustration : Modèle à risques compétitifs pour la mortalité en réanimation

La mortalité des malades admis en réanimation est importante, atteignant en France environ 15% en réanimation et 6 à 25% après la sortie de réanimation [35], soit une mortalité hospitalière d'environ 20 à 40%. Pour mieux établir et comprendre le devenir de ces malades, de nombreux auteurs ont tenté de déterminer les facteurs explicatifs de cette mortalité, c'est à dire rechercher les facteurs qui contribuent à la survenue (ou à la non survenue) du décès, parfois bases de construction de classifications pronostiques [36, 37, 38, 39, 40, 41]. Ces études pronostiques ont ainsi pour objectif de déterminer, parmi les facteurs liés au sujet (notamment en termes de démographie ou de gravité), à la maladie sous jacente, ou aux interventions médicales (par exemple, un traitement antibiotique ou une technique de ventilation assistée), ceux qui sont associés à une sur(sous)-mortalité ou à la survenue d'un événement d'intérêt tel qu'une infection nosocomiale, l'arrêt de la ventilation mécanique ou le sevrage en inotrope.

Si, le plus souvent, l'étude de la mortalité à 28 jours et la modélisation de sa probabilité par un modèle logistique sont privilégiées, l'intérêt se porte parfois sur le délai de survenue du décès ou de tout autre événement d'intérêt. L'utilisation de modèles adaptés aux données de survie parait alors souhaitable. Longtemps, les techniques statistiques proposées en réanimation ont utilisé des approches classiques telles que l'estimateur de Kaplan-Meier et le modèle de Cox en considérant les observations des malades sortis vivants comme des observations censurées à droite non informatives.

Ces données sont cependant un exemple d'une situation de risques compétitifs dans la mesure où chaque sujet est exposé simultanément à deux événements exclusifs, le décès en réanimation et la sortie vivant de réanimation, la survenue de l'un modifiant la probabilité de survenue de l'autre. L'utilisation d'une fonction d'incidence cumulée pour décrire la probabilité de survie en réanimation au cours du temps est alors adaptée. Fonction de sous-répartition, elle permet ainsi de décrire au mieux le fait que lorsque le temps tend vers l'infini, la proportion de sujets admis en réanimation qui décèdent ne tend pas vers 1.

L'estimateur des fonctions d'incidence cumulée spécifiques proposé par Gray [31] et le modèle à risques compétitifs de Fine et Gray [12] sont alors les méthodes adaptées à l'inférence sur la mortalité. Nous avons illustré une telle approche dans un article publié en 2006 dans *Critical Care*.

3.2.1 *Evaluating mortality in intensive care units : contribution of competing risks analyses*

Research

Open Access

Evaluating mortality in intensive care units: contribution of competing risks analyses

Matthieu Resche-Rigon[1], Elie Azoulay[2] and Sylvie Chevret[3]

[1]Medical Doctor, Biostatistics Department, Saint Louis Teaching Hospital-Assistance Publique-Hôpitaux de Paris, 1 avenue Claude Vellefaux, Paris, 75010, France
[2]Medical Doctor, Medical Intensive Care Unit, Saint Louis Teaching Hospital-Assistance Publique-Hôpitaux de Paris, 1 avenue Claude Vellefaux, Paris, 75010, France
[3]Medical Doctor, Biostatistics Department, Saint Louis Teaching Hospital-Assistance Publique-Hôpitaux de Paris, 1 avenue Claude Vellefaux, Paris, 75010, France

Corresponding author: Elie Azoulay, elie.azoulay@sls.ap-hop-paris.fr

Received: 20 May 2005 Revisions requested: 27 May 2005 Revisions received: 8 Sep 2005 Accepted: 27 Oct 2005 Published: 1 Dec 2005

Critical Care 2006, **10**:R5 (doi:10.1186/cc3921)
This article is online at: http://ccforum.com/content/10/1/R5

Abstract

Introduction Kaplan–Meier curves and logistic models are widely used to describe and explain the variability of survival in intensive care unit (ICU) patients. The Kaplan–Meier approach considers that patients discharged alive from hospital are 'non-informatively' censored (for instance, representative of all other individuals who have survived to that time but are still in hospital); this is probably wrong. Logistic models are adapted to this so-called 'competing risks' setting but fail to take into account censoring and differences in exposure time. To address these issues, we exemplified the usefulness of standard competing risks methods; namely, cumulative incidence function (CIF) curves and the Fine and Gray model.

Methods We studied 203 mechanically ventilated cancer patients with acute respiratory failure consecutively admitted over a five-year period to a teaching hospital medical ICU. Among these patients, 97 died before hospital discharge. After estimating the CIF of hospital death, we used Fine and Gray models and logistic models to explain variability hospital mortality.

Results The CIF of hospital death was 35.5% on day 14 and was 47.8% on day 60 (97/203); there were no further deaths. Univariate models, either the Fine and Gray model or the logistic model, selected the same eight variables as carrying independent information on hospital mortality at the 5% level. Results of multivariate were close, with four variables selected by both models: autologous stem cell transplantation, absence of congestive heart failure, neurological impairment, and acute respiratory distress syndrome. Two additional variables, clinically documented pneumonia and the logistic organ dysfunction, were selected by the Fine and Gray model.

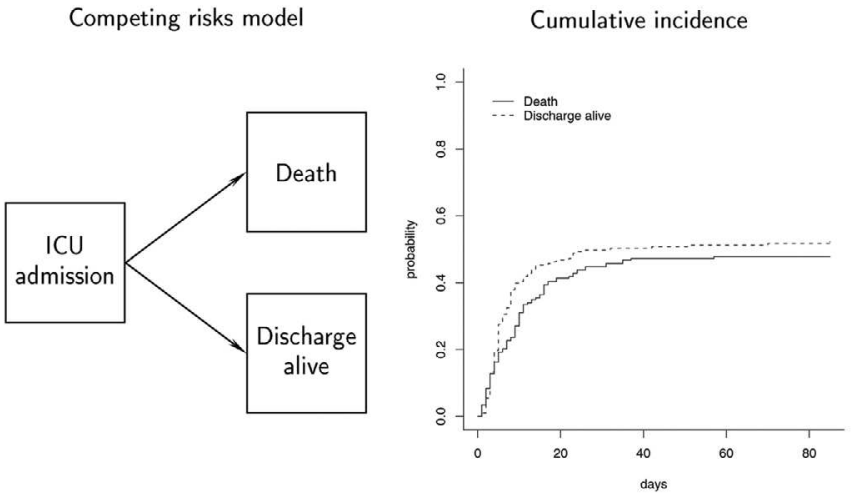
Conclusion The Fine and Gray model appears of interest when predicting mortality in ICU patients. It is closely related to the logistic model, through direct modeling of times to death, and can be easily extended to model non-fatal outcomes.

Introduction

Mortality in intensive care unit (ICU) patients remains high. The estimated mean in France is about 15% for ICU mortality and 6–25% for hospital mortality after ICU discharge [1], yielding a hospital mortality rate of 20–30%, with substantial variations across studies. Reported factors associated with ICU mortality are partly conflicting. Differences in the statistical methods used to estimate mortality and to identify prognostic factors may contribute to these discrepancies. For instance, the outcome of interest could be hospital mortality, ICU mortality, or mortality at a specific time point (e.g. 14 days, 30 days, 60 days, or three months after ICU admission). Furthermore, some studies determine the prevalence of death and others determine the incidence of death. Prevalence of death is estimated from crude mortality ratios (number of deaths divided by number of ICU admissions), and then logistic analysis is used to identify prognostic factors. Incidence studies estimate survival using the Kaplan–Meier method and then look for prognostic factors in Cox models. Many studies use both approaches, determining survival time distributions by the

ARDS = adult respiratory distress syndrome; CIF = cumulative incidence function; ICU = intensive care unit; LOD = logistic organ dysfunction; SCT = stem cell transplantation.

Figure 1



Modelisation of ICU data in the setting of competing risks. Competing risks model (left). Cumulative incidence function of hospital death and being discharged alive (right). ICU, intensive care unit.

Table 1

Hospital mortality according to the main characteristics of the population at admission to the intensive care unit

Characteristic	Hospital mortality
Overall population	97/203 (47.8%)
Autologous stem cell transplantation	19/29 (65.5%)
Clinically documented lung disease	17/27 (63.0%)
Absence of congestive heart failure	3/25 (12.0%)
Neurological impairment	36/52 (61.2%)
Neutropenia	41/71 (57.8%)
Unknown cause of acute respiratory failure	24/42 (57.1%)
Acute respiratory distress syndrome	29/40 (72.5%)

Kaplan–Meier method and then looking for prognostic factors using logistic models [2].

The main argument against using survival methods to analyze ICU mortality or hospital mortality pertains to censoring. Indeed, the Kaplan–Meier method and Cox model assume that censoring is non-informative (for instance, that the survival time of an individual patient is independent of censoring). In other words, patients discharged alive from the hospital must be representative of all other individuals who have survived to this time of discharge but who are still in hospital. In this case, the distribution of the censoring time is unrelated to the distribution of the survival time, so that censoring is 'non-informative' about the mortality pattern of the population. This is likely to be true if the censoring process operates randomly, which is usually the case when mortality is assessed at a point in calendar time (for instance, on 1 January 2005), provided that this time point is selected before the study is initiated.

However this assumption cannot be made if, for example, the survival time of an individual is censored, being withdrawn as a result of a deterioration or an amelioration in his/her physical condition. This is probably the case in the ICU where patients are discharged alive, and thus withdrawn from the survival analysis, because they need no more intensive care, usually due to amelioration or deterioration of their vital conditions. Patients are therefore discharged alive (censored) because they have a lower risk or higher risk of hospital death than the average. These patients are therefore not the same patient population as those who stayed within the hospital. Resulting censoring is 'informative', meaning that censoring carries information about or depends on the survival time. In other words, informative censoring defined a competing risk, given that discharge from the hospital affects the probability of experiencing the event of interest (death before discharge) (Figure 1). In this setting, standard survival methods are no longer valid, and specific methods need to be considered.

Logistic regression, which is widely used to model hospital mortality, is well suited to the described setting. Nevertheless, logistic modeling has been reported to cause loss of information, because it ignores the time to death and the length of hospital stay [3,4]. Specific statistical approaches dedicated to competing risk data, which allow handling of both censoring and time to events, have been proposed [5-7]. They have been applied to ICU data for predicting the occurrence of a non-fatal event in the face of competing mortality [8]. In studies of mortality in ICU patients, since being discharged alive is also a competing risk, these approaches could also apply directly.

Table 2

Univariate prognostic analyses based on logistic regression and Fine and Gray regression

Variable	Logistic model		Fine and Gray model			
	Mortality		Mortality		Discharged alive	
	Odds ratio (95% CI)	*P* value	SHR (95% CI)	*P* value	SHR (95% CI)	*P* value
Autologous stem cell transplantation	2.34 (1.03–5.32)	0.043	1.73 (1.07–2.80)	0.025	0.55 (0.29–1.07)	0.077
Clinically documented lung disease	2.30 (1.09–4.87)	0.029	1. 91 (1.06–3.45)	0.032	0.63 50.42–0.95)	0.027
Absence of congestive heart failure	0.12 (0.04–0.42)	<0.001	0.16 (0.06–0.49)	0.001	2.98 (1.95–4.55)	<0.001
Neurological impairment	3.32 (1.69–6.50)	<0.001	2.35 (1.56–3.55)	<0.001	0.38 (0.23–0.61)	<0.001
Neutropenia	1.85 (1.03–3.33)	0.038	1.61 (1.08–2.39)	0.020	0.65 (0.43–0.97)	0.037
Logistic organ dysfunction	1.20 (1.09–1.33)	<0.001	1.16 (1.09–1.24)	<0.001	0.87 (0.81–0.93)	<0.001
Unknown diagnosis	2.19 (1.13–4.26)	0.021	1.86 (1.20–2.87)	0.005	0.59 (0.35–0.97)	0.039
Acute respiratory distress syndrome	3.68 (1.72–7.89)	<0.001	2.08 (1.39–3.09)	<0.001	0.33 (0.19–0.59)	<0.001

CI, confidence interval; SHR, sub-hazard ratio.

This paper was designed to illustrate the use of competing risks approaches for evaluating the prognostic factors on hospital mortality. To this end, we used a sample of 203 cancer patients admitted to an ICU for acute respiratory failure [9].

Patients and methods

Between 1 November 1997 and 31 October 2002, all adult cancer patients (= 18 years old) admitted to the medical ICU of the Saint-Louis Teaching Hospital, Paris, France for acute respiratory failure were included. Patients in the cohort were followed up until hospital discharge or death. This study has been previously published elsewhere [9] and will now be briefly summarized.

Hospital mortality was the primary endpoint. Standardized forms were used at ICU admission to collect the following: history of autologous or allogeneic stem cell transplantation (SCT), clinically documented lung disease, microbiologically documented invasive aspergillosis, unknown cause of acute respiratory failure, neurological impairment, alveolar hemorrhage, absence of congestive heart failure, acute respiratory distress syndrome (ARDS), neutropenia, logistic organ dysfunction (LOD) score [10], and history of corticosteroid therapy.

Statistical analysis

All statistical analyses were carried out using the SAS 8.2 software package (SAS Inc, Cary, NC, USA) and the R 2.0.1 software package [11].

To describe hospital mortality, we utilized a competing risks model (Figure 1). First, we computed the cumulative incidence function (CIF) of death over time. At time *t*, the CIF defines the probability of dying in the hospital by that time *t* when the population can be discharged alive. Note that, contrarily to a distribution function that tends to 1, the CIF tends to the raw proportion of deaths, so it is also called a 'subdistribution function'. The CIF has been estimated from the data using the *cmprsk* package developed by Gray [12].

To estimate the influence of baseline covariates on hospital mortality, we then used logistic models that estimated the strength of the association between each variable and death based on the odds ratio. Finally, we used the Fine and Gray model [7], which extends the Cox model to competing risks data by considering the subdistribution hazard (for instance, the hazard function associated with the CIF). The strength of the association between each variable and the outcome was assessed using the sub-hazard ratio, which is the ratio of hazards associated with the CIF in the presence of and in the absence of a prognostic factor. In both logistic modeling and Fine and Gray modeling, prognostic factors were evaluated in univariate and multivariate analyses. Models were fitted using the *lrm* and *crr* routines in the R software package [11], respectively. Variables associated with the primary endpoint (hospital death) at the 10% level on the basis of univariate models were introduced in the multivariate models.

Results

Of the 203 patients included in the study, 97 patients (47.8%) died in the hospital. The estimated CIF of death was 35.5% on day 14 and was 47.8% on day 60; no additional deaths occurred after day 60 (Figure 1). Table 1 presents hospital mortality according to the main characteristics at ICU admission. Based on univariate logistic models, eight variables were associated with hospital mortality: autologous SCT, clinically documented lung disease, absence of congestive heart failure, neurological impairment, neutropenia, LOD, unknown cause of acute respiratory failure, and ARDS (Table 2). In the multivariate logistic model, four of these variables supplied independent prognostic information at the 5% level: autolo-

Table 3

The logistic and the Fine and Gray multivariate regression models including eight variables selected on the basis of univariate analyses

Variable	Logistic model		Fine and Gray model			
	Mortality		Mortality		Discharged alive	
	Odds ratio (95% CI)	*P* value	SHR (95% CI)	*P* value	SHR (95% CI)	*P* value
Autologous stem cell transplantation	3.51 (1.37–9.02)	0.009	1.77 (1.00–3.14)	0.049	0.46 (0.23–0.95)	0.035
Clinically documented lung disease	2.01 (0.79–5.14)	0.143	2.09 (1.05–4.15)	0.036	0.71 (0.47–1.08)	0.110
Absence of congestive heart failure	0.12 (0.03–0.57)	0.008	0.22 (0.07–0.64)	0.006	2.20 (1.42–3.42)	<0.001
Neurological impairment	2.63 (1.19–5.81)	0.017	1.84 (1.16–2.91)	0.009	0.44 (0.27–0.71)	<0.001
Neutropenia	1.15 (0.53–2.51)	0.721	1.22 (0.73–2.91)	0.450	0.81 (0.50–1.31)	0.390
Logistic organ dysfunction	1.11 (0.97–1.27)	0.133	1.10 (1.00–1.20)	0.040	0.92 (0.85–1.01)	0.065
Unknown diagnosis	1.82 (0.85–3.89)	0.122	1.20 (0.72–2.00)	0.480	0.71 (0.40–1.27)	0.250
Acute respiratory distress syndrome	3.26 (1.42–7.49)	0.005	1.85 (1.21–2.85)	0.005	0.33 (0.19–0.58)	<0.001

CI, confidence interval; SHR, sub-hazard ratio.

gous SCT, absence of congestive heart failure, neurological impairment, and ARDS (Table 3).

From univariate Fine and Gray models, the same eight variables were associated with the cumulative incidence of hospital death (Table 2). When a multivariate Fine and Gray regression model was used, six of the eight variables were found to supply independent prognostic information at the 5% level: autologous SCT, clinically documented lung disease, absence of congestive heart failure, neurological impairment, LOD, and ARDS (Table 3). Variables associated with the cumulative incidence of being discharged alive from the hospital were the same as those associated with the cumulative incidence of hospital death, except for autologous SCT in univariate analyses (Table 2) and LOD in multivariate analysis (Table 3).

Discussion

Competing risks methods have been used in ICU studies to study non-fatal endpoints in the face of competing mortality [13,14]. We have shown that even mortality in ICU patients can be analyzed using these methods where discharges alive compete with hospital deaths. We illustrated the use and interest of the two main statistical models available in this setting; namely, the logistic model and the Fine and Gray model.

Standard survival analyses are not satisfactory for describing ICU-patient mortality over time: the assumption that censoring is independent of the event of interest (death) is violated, since patients discharged alive are not representative of all other patients still in the hospital. The misuse of the Kaplan–Meier method in this setting is well known and the CIFs have been reported as the optimal tools to measure the probability of the outcome of interest over time [5,6]. The logistic model is widely used, but it ignores the exposure times. Moreover, since the logistic regression does not allow the inclusion of time-dependent covariates [15], one could not adjust for exposure time in that model.

A new regression model, based on CIF-associated hazards, has been proposed by Fine and Gray [7] for identifying prognostic factors in this competing risks setting. This model was first used in cancer patients, to predict non-fatal events such as relapses or metastasis, with death prior to these events as a competing risk [16-21]. We proposed to illustrate the use of the Fine and Gray model for explaining hospital mortality in ICU patients, using a specific R routine [11]. Of note, since all patients either died or were discharged alive by the end of follow-up, standard routines from the SAS package would have been used after recoding exposure times of patients who were discharged alive at the largest observed time of death [5].

Actually, the Fine and Gray model is closely related to the logistic model. The logistic regression focuses attention on the prevalence of hospital death as a measure of prognosis in ICU patients. The Fine and Gray model is based on the hazard associated with the CIF, and therefore predicts the cumulative incidence of death, which tends over time to the prevalence of death. It models this hazard over time, incorporating the different exposure times in the ICU (or hospital) ignored by the logistic model.

As a result, the logistic model and the Fine and Gray model differ in terms of measures used to evaluate the strength of association between the prognostic factor and the hospital death. The sub-hazard ratio on which the Fine and Gray model relies could appear difficult to interpret since it relies on the ratio of subdistribution hazards, which are not directly interpretable in terms of probabilities. Nevertheless, the odds ratio estimated from the logistic model is often misinterpreted as a relative risk – although it should not be misinterpreted unless the outcome

of interest is rare (e.g. <20–30%) [22]. When analyzing the mortality of specific ICU patients, such as the cancer patients of our series, this is clearly untrue. In that sense, the sub-hazard ratio appears to be a better approximation of the relative risk than the odds ratio [23].

Estimates of the separate effects of covariates on each outcome could be provided by both models. In ICU patients, when only two risks are considered and no patients are lost to follow-up, fitting a logistic model using either death or discharge alive will result in mathematically equivalent models. By contrast, the Fine and Gray model considers the hazard of death over time so that distinct effects, although inter-related, could be estimated, reaching distinct *P* values. Indeed, a deleterious effect on one risk is necessarily associated with a protective effect on the other risk: therefore, increased mortality in a patient subset is necessarily associated with a decreased incidence of being discharged alive. This was illustrated in our series. For instance, in patients with autologous SCT, the cumulative incidence of hospital death was significantly increased and that of being discharged alive was decreased, but not significantly (Table 2).

In this paper we have raised the differences between both the logistic model and the Fine and Gray model. Of note, both models present limitations mostly due to the required underlying assumptions (proportional hazards for Fine and Gray; log-linearity and additive effects of covariates for both models).

Finally, we evaluated the competing risks approach for predicting hospital mortality because being discharged alive competes with the outcome of interest. The same method could be used to predict ICU mortality. As mentioned earlier, competing risks analyses based on either the logistic model or the Fine and Gray model may also be valuable for modeling non-fatal outcomes in ICU patients, such as mechanical ventilation or nosocomial infection, with deaths before the outcome of interest and being discharged alive as competing risks [13].

Conclusion

When modeling the mortality of ICU patients, we showed that discharge alive defines a competing risks outcome for hospital (or ICU) mortality. Therefore, besides the widely used logistic regression analyses, standard methods for analyzing competing risks data can be used. Although closely related, the models mostly differ in the handling of exposure times.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MRR and SC conceived the statistical study and drafted the manuscript. MRR performed the statistical analyses. EA conceived and helped to design and coordinate the clinical study. All the authors read and approved the final manuscript.

Key messages

- When estimating the mortality of ICU patients, being discharged alive from the ICU or from the hospital should be considered a competing risk.
- Specific statistical approaches for analyzing outcomes competing against other events are of prime interest in the ICU setting.
- To depict mortality over time, the CIF should be used.
- To predict mortality, the Fine and Gray model, which is based on the subdistribution hazard associated with the CIF, is an alternate to the widely used logistic regression model.

References

1. Azoulay E, Adrie C, De Lassence A, Pochard F, Moreau D, Thiery G, Cheval C, Moine P, Garrouste-Orgeas M, Alberti C, *et al.*: **Determinants of postintensive care unit mortality: a prospective multicenter study.** *Crit Care Med* 2003, **31:**428-432.
2. Vanhems P, Lepape A, Savey A, Jambou P, Fabry J: **Nosocomial pulmonary infection by antimicrobial-resistant bacteria of patients hospitalized in intensive care units: risk factors and survival.** *J Hosp Infect* 2000, **45:**98-106.
3. Azoulay E, Alberti C, Bornstain C, Leleu G, Moreau D, Recher C, Chevret S, Le Gall JR, Brochard L, Schlemmer B: **Improved survival in cancer patients requiring mechanical ventilatory support: impact of noninvasive mechanical ventilatory support.** *Crit Care Med* 2001, **29:**519-525.
4. de Irala-Estevez J, Martinez-Concha D, Diaz-Molina C, Masa-Calles J, Serrano del Castillo A, Fernandez-Crehuet Navajas R: **Comparison of different methodological approaches to identify risk factors of nosocomial infection in intensive care units.** *Intensive Care Med* 2001, **27:**1254-1262.
5. Andersen PK, Abildstrom SZ, Rosthoj S: **Competing risks as a multi-state model.** *Stat Methods Med Res* 2002, **11:**203-215.
6. Pepe M, Mori M: **Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data?** *Stat Med* 1993, **12:**737-751.
7. Fine J, Gray R: **A proportional hazards model for the subditribution of a competing risk.** *J Am Statist Assoc* 1999, **94:**496-509.
8. Rubenfeld G: **Looking beyond 28-day all-cause mortality.** *Crit Care* 2002, **6:**293-294.
9. Azoulay E, Thiéry G, Chevret S, Moreau D, Darmon M, Bergeron A, Yang K, Meignin V, Ciroldi M, Le Gall J, *et al.*: **The prognosis of acute respiratory failure in critically ill cancer patients.** *Medicine (Baltimore)* 2004, **83:**360-70.
10. Le Gall J, Klar J, Lemeshow S, Saulnier F, Alberti C, Artigas A, Teres D: **The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group.** *JAMA* 1996, **276:**802-810.
11. R Foundation for Statistical Computing: **R: a language and environment for statistical computing (version 2.0.1).** [http://www.R-project.org].
12. ***cmprsk* package** [http://biowww.dfci.harvard.edu/~gray/cmprsk_2.1-4.tar.gz]
13. Alberti C, Metivier F, Landais P, Thervet E, Legendre C, Chevret S: **Improving estimates of event incidence over time in populations exposed to other events: application to three large databases.** *J Clin Epidemiol* 2003, **56:**536-545.
14. Alberti C, Brun-Buisson C, Chevret S, Antonelli M, Goodman SV, Martin C, Moreno R, Ochagavia AR, Palazzo M, Werdan K, *et al.*: **Systemic inflammatory response and progression to severe sepsis in critically ill infected patients.** *Am J Respir Crit Care Med* 2005, **171:**461-468.
15. Leffondre K, Abrahamowicz M, Siemiatycki J: **Evaluation of Cox's model and logistic regression for matched case-control data**

with time-dependent covariates: a simulation study. *Stat Med* 2003, **22:**3781-3794.
16. Castaigne S, Chevret S, Archimbaud E, Fenaux P, Bordessoule D, Tilly H, de Revel T, Simon M, Dupriez B, Renoux M, *et al.*: **Randomized comparison of double induction and timed-sequential induction to a '3+7' induction in adults with acute myeloid leukemia (AML). Long-term analysis of the Acute Leukemia French Association (ALFA) 9000 study.** *Blood* 2004, **104:**2467-74.
17. Colleoni M, O'Neill A, Goldhirsch A, Gelber RD, Bonetti M, Thürlimann B, Price KN, Castiglione-Gertsch M, Coates AS, Lindtner J, *et al.*: **Identifying breast cancer patients at high risk for bone metastases.** *J Clin Oncol* 2000, **18:**3925-3935.
18. de Botton S, Coiteux V, Chevret S, Rayon C, Vilmer E, Sanz M, de La Serna J, Philippe N, Baruchel A, Leverger G, *et al.*: **Outcome of childhood acute promyelocytic leukemia with all-trans-retinoic acid and chemotherapy.** *J Clin Oncol* 2004, **22:**1404-1412.
19. Martelli G, Miceli R, De Palo G, Coradini D, Salvadori B, Zucali R, Galante E, Marubini E: **Is axillary lymph node dissection necessary in elderly patients with breast carcinoma who have a clinically uninvolved axilla?** *Cancer* 2003, **97:**1156-1163.
20. Robson ME, Chappuis PO, Satagopan J, Wong N, Boyd J, Goffin JR, Hudis C, Roberge D, Norton L, Bégin LR, *et al.*: **A combined analysis of outcome following breast cancer: differences in survival based on BRCA1/BRCA2 mutation status and administration of adjuvant cancer.** *Breast Cancer Res* 2004, **6:**8-17.
21. Shpall EJ, Quinones R, Giller R, Zeng C, Baron AE, Jones RB, Bearman SI, Nieto Y, Freed B, Madinger N, *et al.*: **Transplantation of Ex vivo expanded cord blood.** *Biol Blood Marrow Transplant* 2002, **8:**368-376.
22. Altman DG, Deeks JJ, Sackett DL: **Odds ratios should be avoided when events are common.** *BMJ* 1998, **317:**1318.
23. Symons MJ, Moore DT: **Hazard rate ratio and prospective epidemiological studies.** *J Clin Epidemiol* 2002, **55:**893-899.

3.3 Mesure d'influence pour la fonction de risque cause-spécifique

Par souci de simplification, les coefficients de régression β ne sont plus indicés par modèle.

Dans le cadre du modèle Cox, des méthodes de mesure de l'influence individuelle ont été initialement développées à partir des résidus. Les premiers travaux concernant les résidus furent réalisés à la fin des années 1970 et au début des années 1980, essentiellement par Schoenfeld [42]. Considérons n observations (T_i^*, δ_i). Pour le modèle de Cox décrit par l'équation (3.5), les résidus de Schoenfeld se définissent par :

$$r_i = \delta_i \left(X_i - \frac{\sum_{j \in R_i} X_j \exp(X_j^T \beta)}{\sum_{j \in R_i} \exp(X_j^T \beta)} \right), \ i = 1, ..., n \tag{3.17}$$

En 1994, Grambsch et Therneau [43] proposèrent une version des résidus de Schoenfeld réduits par la matrice de variance-covariance du modèle et pondérés par le nombre de décès observés parmi les n individus. Cette version est considérée comme plus sensible pour la détection d'observations aberrantes [10] .

En 1984, Cain et Lange proposèrent d'écrire une valeur approchée des $dbeta = \hat{\beta} - \hat{\beta}_{(i)}$ à partir d'un développement limité d'ordre 1 [9]. Comme pour le modèle linéaire et les modèles linèaires généralisés, ils utilisèrent une fonction de log-vraisemblance partielle pondérée :

$$L(\beta, w) = \sum_{i=1}^{n} \delta_i w_i \left(X_i^T \beta - \log \left(\sum_{j \in R_i} w_j \exp\left(X_j^T \beta\right) \right) \right) \tag{3.18}$$

où $w = (w_i, ..., w_n)$ est un n vecteur de poids, et $R_i = \{j : (T_j^* \geq T_i)\}$.

En utilisant un développement limité d'ordre 1 en w_0 le vecteur unité de dimension n, ils écrivent $\hat{\beta} - \hat{\beta(w)} \simeq \left[\frac{\partial \hat{\beta(w)}}{\partial w}\right]_{\hat{\beta}, w_0} (w - w_0)$, puis $\frac{\partial \hat{\beta(w)}}{\partial w} = I^{-1} \frac{\partial U(\beta, w)}{\partial w}$, avec U le vecteur score et I la matrice d'information de Fisher. Ils développent surtout un algorithme de calcul simple et rapide de $\frac{\partial U(\beta,w)}{\partial w}$, et formalisent le lien qui existe entre $\frac{\partial U(\beta,w)}{\partial w}$ et les résidus de Schoenfeld. Ils montrent en effet que $\frac{\partial U(\beta,w)}{\partial w_i}$, qui représente l'influence de l'individu i sur le vecteur score, peut s'écrire sous la forme :

$$\left[\frac{\partial U(\beta, w)}{\partial w_i}\right]_{\hat{\beta}, w_0} = \delta_i \{X_i - \hat{E}(X|R_i)\} - \sum_{l \in D_i} \frac{\exp\left(X_i^T \hat{\beta}\right)}{\sum_{j \in R_l} \exp\left(X_j^T \hat{\beta}\right)} \{X_i - \hat{E}(X|R_l)\} \tag{3.19}$$

avec $\hat{E}(X|R_i) = \sum_{j \in R_i} X_j \exp\left(X_j^T \hat{\beta}\right) / \sum_{j \in R_i} \exp\left(X_j^T \hat{\beta}\right)$ et D_i l'ensemble des individus décédés avant T_i. On reconnait alors $\delta_i \{X_i - \hat{E}(X|R_i)\}$ comme l'expression du résidu de Schoenfeld pour l'individu i.

En 1985, Reid et Crepeau [44], clarifient la notion de fonction d'influence pour le modèle de Cox. Enfin, en 1985, Storer et Crowley [45] proposent un calcul des *dfbetas* basés un modèle de régression augmentée et une approximation faite par une itération de la procédure de Raphson-Newton.

A partir de ces calculs de $(\hat{\beta} - \hat{\beta_{(i)}})$, l'estimation de la distance de Cook s'effectue en utilisant la formule (2.7) :

$$D_i = \frac{1}{p}(\hat{\beta} - \hat{\beta_{(i)}})^T I(\hat{\beta} - \hat{\beta_{(i)}}) \tag{3.20}$$

A la suite de la définition de l'influence locale individuelle de Cook en 1986 [3], Pettitt et Bin Daud [23] étendent, en reprenant la vraisemblance partielle pondérée écrite par Cain et Lange [9], l'influence locale au modèle de Cox. La vraisemblance pondérée s'écrit alors :

$$L(\beta, w) = \sum_{i=1}^{n} \mathbf{1}_{[\varepsilon_i=1]} w_i \left(X_i^T \beta - \log \left(\sum_{j \in R_i} w_j \exp\left(X_j^T \beta\right) \right) \right) \tag{3.21}$$

et la mesure d'influence locale est l_{max} le vecteur propre de la plus grande valeur propre de la matrice $\Delta^T I^{-1} \Delta$. Barlow [46], en 1997 propose de remplacer I par l'inverse de la matrice de covariance robuste, mais cette approche ne semble pas avoir été reprise par la suite. Il applique simplement une norme légèrement différente aux $(\hat{\beta} - \hat{\beta_{(i)}})$ proposée en 1982 par Cook [2] dans la cadre du modèle linéaire et l'étend à l'influence locale.

En 1992, Escobar et Meeker [5] proposent une revue des méthodes de mesures d'influence pour les modèles de survie et une extension aux modèles de temps de survie accélérés. Comme Collett par la suite [10], ils utilisent la mesure d'influence locale l_{max} pour déterminer les individus influençants.

Par la suite, plusieurs auteurs proposeront des adaptations de ces concepts au modèle de Cox en temps discrets [47], et au modèle de Cox à variables dépendantes du temps [48].

En 1999, Wei et Su [49] proposent d'utiliser les mesures d'influence locale dans le cadre d'un choix entre plusieurs modèles de survie, privilégiant le modèle le plus stable. Enfin Wei et Kosorok [50], partant des travaux de Lawrance sur l'influence masquée [19] définirent, en 2000, une mesure d'influence pour les modèles à risques proportionnels basés sur l'exclusion itérative de deux observations. Ils définissent ainsi, comme Cook, une norme pour $\hat{\beta} - \hat{\beta}_{(i)} - \hat{\beta}_{(k)} + \hat{\beta}_{(i,k)}$ à l'aide de la matrice d'information de Fisher, sans pour autant diviser la norme par p le nombre de paramètres du modèle :

$$\left\| \hat{\beta} - \hat{\beta}_{(i)} - \hat{\beta}_{(k)} + \hat{\beta}_{(i,k)} \right\| = (\hat{\beta} - \hat{\beta}_{(i)} - \hat{\beta}_{(k)} + \hat{\beta}_{(i,k)})^T I(\hat{\beta} - \hat{\beta}_{(i)} - \hat{\beta}_{(k)} + \hat{\beta}_{(i,k)}) \tag{3.22}$$

3.4 Mesure d'influence pour le modèle de Fine et Gray

3.4.1 Développement d'une mesure d'influence locale

Partant essentiellement des travaux de Cain et Lange [9] et Pettitt et Bin Daud [23], nous avons proposé d'étendre la mesure d'influence locale au modèle de Fine et Gray [12]. Cette extension permet, par ailleurs, un calcul simple de la distance de Cook à l'aide des $(\hat{\beta} - \hat{\beta_{(i)}})$.

Considérons une situation où un individu est exposé à k risques en compétition. Soient n observations (T_i^*, ε_i^*), $\varepsilon \in \{1, ...k\}$. Chaque individu i, $i = 1, ..., n$, est caractérisé par un p-vecteur de covariables X_i. Soit $w = (w_i, ..., w_n)$ un n vecteur de poids.

La fonction de log-vraisemblance partielle pondérée pour l'événement de type 1 dans le modèle de Fine et Gray s'écrit :

$$L(\beta, w) = \sum_{i=1}^{n} \mathbf{1}_{[\varepsilon_i=1]} w_i \left(X_i^T \beta - \log \left(\sum_{j \in R_i} v_{ji} w_j \exp\left(X_j^T \beta\right) \right) \right) \tag{3.23}$$

où $R_i = \{j : (T_j \geq T_i) \cup (T_j \leq T_i \cap \varepsilon_j \notin \{0, 1\})\}$, $v_{ji} = \hat{G}(T_i)/\hat{G}(T_j \wedge T_i)$ définit l'ensemble des individus à risque, et $\hat{G}$ est l'estimation par la méthode de Kaplan-Meier de la fonction de survie des délais de censure. Les individus n'expérimentant pas l'événement d'intérêt, sont donc inclus dans tous les R_i.

Soit $w_0 = (1, ..., 1)$ le vecteur unité de taille n. Le principe est de mesurer l'influence des individus en étudiant les variations dans l'évaluation des coefficients de regression engendrées par de petites perturbations autour de w_0. Considérons $L(\beta, w)$ comme deux fois différentiable par rapport à β et w. Soit $\hat{\beta}$ et $\hat{\beta(w)}$ les estimations du maximum de vraisemblance respectivement de $L(\beta, w_0)$ et de $L(\beta, w)$. Le déplacement de vraisemblance (LD) se définit comme précédemment par :

$$LD(w) = 2\left[L(\hat{\beta}) - L(\hat{\beta(w)})\right] \tag{3.24}$$

Au voisinage de w_0, une approximation par un développement limité d'ordre 2 permet d'écrire :

$$LD(w) \approx [\hat{\beta} - \hat{\beta(w)}]^T I [\hat{\beta} - \hat{\beta(w)}] \tag{3.25}$$

où I est la matrice d'information de Fisher.

L'approximation d'ordre 1 de $\hat{\beta} - \hat{\beta(w)}$ donne :

$$\hat{\beta} - \hat{\beta(w)} \approx I^{-1} \left[\frac{\partial U(\beta(w))}{\partial w} \right]_{\hat{\beta}, w_0} (w_0 - w) \tag{3.26}$$

où $U(\beta(w))$ est le vecteur score.

On obtient donc :

$$LD(w) \approx (w_0 - w)^T \Delta^T I^{-1} \Delta (w_0 - w) \tag{3.27}$$

où $\Delta = (\Delta_1, \ldots, \Delta_n)$ est une matrice de dimension $p \times n$, dont les Δ_i sont définis par :

$$\Delta_i = \left[\frac{\partial U(\beta(w)}{\partial w_i} \right]_{\hat{\beta}, w_0} \tag{3.28}$$

Soit $\hat{E}(X|R_i)$ définie par $\sum_{j \in R_i} v_{ji} X_j \exp(X_j^T \hat{\beta}) / \sum_{j \in R_i} v_{ji} \exp(X_j^T \hat{\beta})$, avec $D_i : \{l : (T_l \leq T_i) \cup ((T_l \geq T_i) \cap (\varepsilon_i \notin \{1, 0\}))\}$. Remarquons que pour les individus qui expérimentent un autre événement que celui d'intérêt, D_i inclut tous les individus. Comme Cain et Lange [9] l'ont montré pour le modèle de Cox, Δ_i peut s'écrire sous la forme de la somme de deux termes :

$$\Delta_i = \Delta_{i,1} + \Delta_{i,2} \tag{3.29}$$

où $\Delta_{i,1} = 1_{[\varepsilon_i = 1]}(X_i - \hat{E}(X|R_i)$ correspond aux résidus de Schoenfeld du modèle de Fine et Gray[42] et $\Delta_{i,2}$ est défini par :

$$\Delta_{i,2} = -\sum_{l \in D_i} \mathbf{1}_{[\varepsilon_l = 1]} \frac{v_{il} \exp(X_i^T \hat{\beta})}{\sum_{j \in R_l} v_{jl} \exp(X_j^T \hat{\beta})} \{Z_i - \hat{E}(X|R_l)\} \tag{3.30}$$

On note que $\Delta_{i,2}$ dépend de la présence de l'individu i dans l'ensemble à risque des autres individus.

De la même manière que Pettitt et Bin Daud [23] définissent l'influence locale pour le modèle de Cox, l'influence locale du modèle de Fine et Gray est définie par la valeur absolue $|l_{\max}|$ des coordonnées du vecteur propre correspondant à la plus grande valeur propre de $\Delta^T I^{-1} \Delta$. Comme dans le cas du modèle linéaire, une valeur élévée de la courbure indique une sensibilité du modèle aux petites variations. Par ailleurs, pour tout individu i, l'obtention des Δ_i permet le calcul approché des $(\hat{\beta} - \hat{\beta_{(i)}})$ en utilisant :

$$\hat{\beta} - \hat{\beta_{(i)}} \simeq I^{-1} \Delta_i \tag{3.31}$$

Enfin, on calcule la distance de Cook pour chaque individu, en écrivant :

$$\frac{1}{p} (\hat{\beta} - \hat{\beta_{(i)}})^T I (\hat{\beta} - \hat{\beta_{(i)}}) = \frac{1}{p} \Delta_i^T I^{-1} \Delta_i \tag{3.32}$$

3.4.2 Etude du modèle par une étude de simulation

Pour aller plus loin dans la compréhension du modèle de Fine et Gray, suivant l'approche de Wei et Su [49], nous avons utilisé la mesure d'influence locale développée dans la section précente. Une étude de simulation visant à évaluer l'impact du rang de l'observation a été réalisée.

La méthode de simulation des données est celle décrite par Fine et Gray [12], à la différence près que l'on ne considère ici qu'une covariable X binaire. La fonction de sous-répartition du délai de survenue de l'événement de type 1 est définie par :

$$Pr(T \leq t, \varepsilon = 1|X) = 1 - \left[1 - p' \left\{1 - \exp\left(-t\right)\right\}\right]^{\exp\left(\beta X\right)}$$

où $p' = P(\varepsilon = 1|X = 0)$. Pour chaque échantillon simulé de n observations, on tire au sort une valeur de X dans une loi de Bernoulli $B(0.5)$; on tire au sort une variable U dans une loi uniforme sur $[0; 1]$, puis on déduit :

$$T = -\log\left\{1 - \frac{1 - (1 - U)^{-\exp\left(\beta X\right)}}{p'}\right\}$$

où β est fixé et p' est calculé numériquement de façon à contrôler la prévalence de l'événement de type 1 sur l'ensemble de l'échantillon. Si T n'est pas défini ($U < 1-(1-p')^{\exp\left(\beta X\right)}$), alors le sujet est considéré comme ayant réalisé l'événement de type 2 ($\varepsilon = 2$) et son temps d'événement est tiré au sort dans une loi exponentielle unité. Les temps de censure sont générés en utilisant une distribution exponentielle de paramètre 0.25 de façon à obtenir un pourcentage d'observations censurées d'environ 20%. On calcule ensuite l'influence locale. On simule de façon indépendante, N échantillons, et on estime l'influence locale pour chaque rang à partir de la médiane de la distribution obtenue sur l'ensemble des échantillons.

Nous avons alors montré que l'influence locale est liée au rang de survenue de l'événement ou de la censure. L'influence des individus ayant expérimenté l'événement d'intérêt commence par décroître avec leur rang pour ré-augmenter avec les rangs les plus élevés. Comme on pouvait s'y attendre, l'influence des individus ayant expérimenté l'autre événement que celui d'intérêt ne dépend que faiblement du rang. Ils restent, en effet, dans le modèle de Fine et Gray dans tous les ensembles à risque pour tous les temps T_i. Enfin l'influence des individus censurés croit avec le temps. Le développement de cette mesure d'influence a fait l'objet d'une publication dans *Statistics in Medicine* en 2006.

3.4.3 *Local influence for the subdistribution of a competing risk*

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:1937–1947
Published online 12 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2354

Local influence for the subdistribution of a competing risk

Matthieu Resche-Rigon*,† and Sylvie Chevret‡

Département de Biostatistique et Informatique Médicale, Hôpital Saint-Louis, Université Paris 7, Inserm U717, Paris, France

SUMMARY

Local influence measures have been shown to be a useful tool in identifying influential individuals, and assessing model behaviour towards small perturbations. In the setting of competing risks, we developed a measure of local influence for the Fine and Gray model for the subdistribution hazard. The plot of local influence showed some relationship with time failure rank. This was illustrated on a real data set from a randomized clinical trial in acute promyelocytic leukaemia and on a simulation study. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: competing risks; cumulative incidence; influence; subdistribution hazard

INTRODUCTION

Statistical models are important devices for understanding the essential features of a data set, although they approximate more complicated processes [1]. Therefore, sensitivity studies of model assumptions, uncertainties in data, and other inputs, are very important [2, 3]. In this context, influence analyses, that study how relevant perturbations affect specified key results, have been developed for evaluating model misspecification and for detecting influential points.

To detect influential observations in data analyses, influence diagnostics were first based on case-deletion measures such as the Cook's distance and the likelihood distance that have become popular in practice. Indeed, they have been used in the context of linear regression (see review in Reference [4]), then adapted to generalized linear models and non-linear regression, including Cox models [5–7]. However, case deletion is not the only paradigm that has been used to investigate the consequences of perturbations of a statistical model. In contrast to global measures, Cook used the concept of normal curvature to propose local measures of influence based on case-weight perturbation schemes as a general approach for assessing the influence of a minor perturbation to a statistical model involving linear or non-linear estimation

*Correspondence to: Matthieu Resche-Rigon, Département de Biostatistique et Informatique Médicale, Hôpital Saint-Louis, 1 avenue Claude Vellefaux, 75475 Paris cedex 10, France.
†E-mail: matthieu.resche-rigon@paris7.jussieu.fr
‡E-mail: sylvie.chevret@paris7.jussieu.fr

Received 27 August 2004
Accepted 11 July 2005

[1, 2]. More precisely, this approach relies on the likelihood displacement function, assessing the influence of case-weight perturbation by discrepancies between the likelihood functions of the postulated model and the perturbated model for the data. Recently, Zhu and Zhang showed that local measures of influence are closely related to the case-deletion measures, and derived a test statistic for evaluating model misspecification and for detecting influential points simultaneously [8].

Applications of local influence analysis to specific survival models have been published. Notably, Pettitt and Bin Daud applied local influence methods for the Cox proportional hazards model [9], while Escobar and Meeker developed local influence measures for the accelerated failure time regression model involving left-censored and interval-censored data [3]. In the setting of competing risks, i.e. when several exclusive competing risks act on the population, a semi-parametric regression model for the hazard associated with the subdistribution function has been proposed by Fine and Gray [10]. In recent years, these subdistribution functions—also called the cumulative incidence functions (CIF)—have become important in medical research [11]. However, no influence measure was proposed in this setting.

In this paper, we review and extend the application and interpretation of local influence measures, using competing-risks right-censored data with a semi-parametric regression model for the subdistribution hazard as the motivating problem. The paper is organized as follows. Section 2 reviews the Fine and Gray semi-parametric regression model for competing-risks data, with model assumptions. Section 3 considers the use of likelihood displacement to measure influence for this model. In Section 4, we illustrate these ideas using a real data set from a phase III randomized clinical trial conducted in acute leukaemia. Section 5 presents a simulation study to show the importance of assessing influence as a function of failure ranks. Finally, Section 6 provides a discussion.

REGRESSION MODEL FOR THE SUBDISTRIBUTION HAZARD

In the competing-risks setting, subjects may fail from one out of K distinct and exclusive causes of failure. Thus, observed data typically consist of n independent observations $(X_i, \varepsilon_i, Z_i)$, where X_i is the minimum of the failure time (T_i) and the right-censoring time (C_i), $\varepsilon_i \in \{1, \ldots, K\}$ denotes the failure cause and $\varepsilon_i = 0$ denotes a right-censored observation, and Z_i is a row-vector of p regressors. Consider that we are interested in estimating the effect of an exposure, Z, on a particular cause of failure, identified by $\varepsilon = 1$. Two main approaches have been proposed [12]. The most common approach is to focus on the modelling of the cause-specific hazard of this failure cause [13] through the use of a Cox model. The second approach is to compare the CIF, also called the subdistribution functions, of this failure cause between exposed and unexposed groups, either directly [14, 15], or by modelling the hazard function associated with the CIF [10, 16]. The so-called subdistribution hazard is given by

$$\begin{aligned}\alpha_k(t;Z) &= -d\log\{1 - F_k(t;Z)\}/dt \\ &= f_k(t;Z)/(1 - F_k(t;Z)) \\ &= \lim_{\delta t \to 0} \frac{1}{\delta t} \Pr(t \leqslant T \leqslant t + \delta t, \varepsilon = k | T \geqslant t \cup (T \leqslant t \cap \varepsilon \neq k), Z)\end{aligned}$$

where $F_k(t;Z) = \Pr(T \leqslant t, \varepsilon = k/z)$ are the CIF with subdensities $f_k(t;Z)$, $k \in \{1,\ldots,K\}$. A semi-parametric proportional hazards model was proposed by Fine and Gray to test covariate effects on the subdistribution hazard for the failure cause of interest [10], as follows:

$$\alpha_1(t;Z) = \alpha_{0,1}(t)\exp(\beta Z)$$

where $\alpha_{0,1}(t)$ is an arbitrary non-negative function, and β is a column vector of p unknown regression parameters for Z. Of note, it is not required that the underlying processes leading to failures from different causes are acting independently for a given subject [14, 17]. Using the partial likelihood principle and weighted estimated equations, consistent estimators of the covariate effects were derived, either in the absence of censoring or in the presence of independent right censoring [10].

LOCAL INFLUENCE MEASURES

Following Cook [2], let $\omega = (\omega_1,\ldots,\omega_n)^{\mathrm{T}} \in \mathbb{R}^n$ denote a perturbation vector of n observations. Let $L(\beta)$ and $L(\beta|\omega)$ denote the log likelihood functions of the postulated model and the perturbated model, respectively. In the Fine and Gray model, $L(\beta|\omega)$ could be written as follows:

$$L(\beta|\omega) = \sum_{i=1}^{n} \mathbf{1}_{[\varepsilon_i = 1]}\omega_i \left(Z_i^{\mathrm{T}}\beta - \log\left(\sum_{j \in R_i} v_{ji}\omega_j \exp\left(Z_j^{\mathrm{T}}\beta\right)\right)\right)$$

where $\mathbf{1}_{[\,]}$ is the indicator function, R_i is the risk set given by $\{j : (X_j \geqslant X_i) \cup (X_j \leqslant X_i \cap \varepsilon_j \notin \{0,1\})\}$, $v_{ji} = \hat{G}(X_i)/\hat{G}(\min(X_j, X_i))$, and $\hat{G}$ is the Kaplan–Meier estimate of $G(t) = \Pr(C > t)$. Following Fine and Gray [10], note that R_i is an 'unnatural' risk set, since individuals who have already failed from another causes than that of interest stay at risk after their failure time.

Let $\omega_0 = (1,1,\ldots,1)^{\mathrm{T}} \in \mathbb{R}^n$, such that $L(\beta) = L(\beta|\omega_0)$ for all β, corresponding to the null perturbation. We also assume that $L(\beta|\omega)$ is twice differentiable with respect to (β,ω). Let $\hat{\beta}$ and $\hat{\beta}_\omega$ denote the maximum partial likelihood estimators derived from $L(\beta)$ and $L(\beta|\omega)$, respectively. The likelihood displacement function (LD) measures distance between $\hat{\beta}$ and $\hat{\beta}_\omega$ in terms of log likelihood difference, and allows the evaluation of influence that a small perturbation ω has on $\hat{\beta}$, as defined by Cook:

$$\mathrm{LD}(\omega) = 2[L(\hat{\beta}) - L(\hat{\beta}_\omega)] \tag{1}$$

A Taylor series approximation in the neighbourhood of ω_0 yields

$$\mathrm{LD}(\omega) \approx [\hat{\beta} - \hat{\beta}_\omega]^{\mathrm{T}} I [\hat{\beta} - \hat{\beta}_\omega] \tag{2}$$

where I denotes the Fisher information matrix, and $\hat{\beta} - \hat{\beta}_\omega$ can be approximated to the first order as

$$\hat{\beta} - \hat{\beta}_\omega \approx I^{-1}\left[\frac{\partial U(\beta|\omega)}{\partial \omega}\right]_{\hat{\beta},\omega_0} (\omega_0 - \omega)$$

where $U(\beta|\omega)$ denotes the score vector, yielding

$$\mathrm{LD}(\omega) \approx (\omega_0 - \omega)^{\mathrm{T}} \Delta^{\mathrm{T}} I^{-1} \Delta (\omega_0 - \omega)$$

where $\Delta = (\Delta_1, \ldots, \Delta_n)$ is a $p \times n$ matrix, with Δ_i a $p \times 1$ vector defined by

$$\Delta_i = \left[\frac{\partial U(\beta|\omega)}{\partial \omega_i} \right]_{\hat{\beta}, \omega_0}$$

Actually, Δ_i represents the change in the score vector $U(\beta|\omega)$ due to the minor perturbation ω_i of the ith observation. Following Cain and Lange in the setting of the Cox model [5], let $\hat{E}(Z|R_i)$ be $\sum_{j \in R_i} v_{ji} Z_j \exp(Z_j^{\mathrm{T}} \hat{\beta}) / \sum_{j \in R_i} v_{ji} \exp(Z_j^{\mathrm{T}} \hat{\beta})$ and D_i denote the set of individuals whose risk set includes patient i. Thus, D_i are defined as $\{l : (X_l \leqslant X_i) \cup ((X_l \geqslant X_i) \cap (\varepsilon_i \notin \{1, 0\}))\}$. Note that for individuals i who failed from another cause than that of interest, D_i includes all individuals. Δ_i can be defined as the sum of two components:

$$\Delta_i = \Delta_{i,1} + \Delta_{i,2} \tag{3}$$

where $\Delta_{i,1} = \mathbf{1}_{[\varepsilon_i = 1]}(Z_i - \hat{E}(Z|R_i))$ is a Schoenfeld-type residual [18], and

$$\Delta_{i,2} = -\sum_{l \in D_i} \mathbf{1}_{[\varepsilon_l = 1]} \frac{v_{il} \exp(Z_i^{\mathrm{T}} \hat{\beta})}{\sum_{j \in R_l} v_{jl} \exp(Z_j^{\mathrm{T}} \hat{\beta})} \{Z_i - \hat{E}(Z|R_l)\}$$

is related to the presence of the ith individual in the risk set of other individuals.

Cook used concepts from differential geometry to describe the local influence achieved by small perturbations on inferences [2]. He defined the curvature of $\mathrm{LD}(\omega)$ at ω_0 in the direction l as C_l, with $C_{\max}$ corresponding to the maximum eigenvalues of the $n \times n$ matrix $\Delta^{\mathrm{T}} I^{-1} \Delta$. Based on empirical experience, a value of $C_{\max} \geqslant 2$ could indicate a 'notable local sensitivity' [2], though such a limit appears somewhat arbitrary and dependent on the scale of the perturbation parameter, ω. The n-eigenvector $l_{\max}$ associated with $C_{\max}$ can be used to indicate how to perturb the postulated model to obtain the greatest local change in the likelihood displacement, and was proposed as the most important diagnostic to come from this approach [2]. The ith element of $|l_{\max}|$ is considered as a measure of the local influence of the ith observation.

EXAMPLE

The illustration concerns a multicentre phase III randomized clinical trial conducted in patients with acute promyelocytic leukaemia [19]. Patients were first randomized between two induction treatment groups, namely the association of all transretinoic acid (ATRA) plus chemotherapy (CT) *versus* ATRA followed by CT (induction control group). Patients achieving complete remission ($n = 229$) were secondly randomized using factorial design between four maintenance treatment groups, according to further administration of ATRA and CT, or not. Relapse following second randomization defined the endpoint of interest. Initial analysis was based on a cause-specific model, where deaths prior to relapse were censored at time of death. Actually, in this data set, there were 41 relapses and 11 deaths prior to relapse.

Table I. APL93 trial.

Rank of deleted case	Local influence $\lvert l_{\max}\rvert$	Covariable	$\hat{\beta}$	Relative change in $\hat{\beta}$	Estimated SD($\hat{\beta}$)
None		$\text{ATRA} + \text{CT}_{\text{induction}}$	−0.229		0.330
		$\text{ATRA}_{\text{maintenance}}$	−0.638		0.329
		$\text{CT}_{\text{maintenance}}$	−0.468		0.323
8	0.29	$\text{ATRA} + \text{CT}_{\text{induction}}$	−0.162	0.29	0.331
		$\text{ATRA}_{\text{maintenance}}$	**−0.713**	0.12	0.334
		$\text{CT}_{\text{maintenance}}$	−0.527	0.13	0.326
11	0.29	$\text{ATRA} + \text{CT}_{\text{induction}}$	−0.161	0.30	0.331
		$\text{ATRA}_{\text{maintenance}}$	**−0.712**	0.12	0.334
		$\text{CT}_{\text{maintenance}}$	−0.526	0.12	0.326
12	0.29	$\text{ATRA} + \text{CT}_{\text{induction}}$	−0.161	0.30	0.331
		$\text{ATRA}_{\text{maintenance}}$	**−0.712**	0.12	0.334
		$\text{CT}_{\text{maintenance}}$	−0.526	0.12	0.326
218	0.11	$\text{ATRA} + \text{CT}_{\text{induction}}$	−0.200	0.13	0.332
		$\text{ATRA}_{\text{maintenance}}$	**−0.658**	0.03	0.330
		$\text{CT}_{\text{maintenance}}$	−0.496	0.06	0.324
229	0.11	$\text{ATRA} + \text{CT}_{\text{induction}}$	−0.200	0.13	0.332
		$\text{ATRA}_{\text{maintenance}}$	**−0.658**	0.03	0.330
		$\text{CT}_{\text{maintenance}}$	−0.496	0.06	0.324

Note: Estimation of regression coefficients with estimated standard deviation (SD) of the effects of induction ATRA+CT and maintenance therapies with ATRA or CT, based on a multivariable Fine and Gray model, using the whole sample and after exclusion of individual cases. Significant effects at the 0.05 level in bold.

We applied a CIF-based analysis, treating death prior relapse as a competing failure cause of relapse. We considered three explanatory variables, all of which are indicator variables of the randomized groups: The first, z_1, gives the induction randomization group, where $z_1 = 1$ denotes the ATRA+CT group; the second, z_2, denotes the allocation to a maintenance arm with ATRA; and the third, z_3, denotes the allocation to a maintenance arm with CT. Of note, each variable was roughly balanced, with 60.3 per cent cases with $z_1 = 1$, 51.5 per cent cases with $z_2 = 1$ and 64.2 per cent cases with $z_3 = 1$. We first fitted a Fine and Gray model with z_1, z_2 and z_3 as explanatory variables. The estimated regression coefficients are reported in Table I. No effect was significant at the 5 per cent level. Local influences against ranks of failure or censoring times are displayed in Figure 1. The three cases in the upper left corner of Figure 1 have the largest estimated influences upon $\hat{\beta}$. These patients, with rank 8, 11 and 12, all relapsed early and had the three covariate values at 1; by contrast, remaining patients with $z = (1, 1, 1)$ did not fail from either cause and were censored much later. Otherwise, the two cases in the lower right corner appear to be different from the others. Although they were censored, these patients had long observation times, namely of ranks 218 and 229. Table I summarizes the effects of deleting either one of these five cases, then refitting the model. With either one removed, the estimated subdistribution hazard of relapse of the ATRA maintenance arm was decreased from 0.53 (95 per cent confidence interval:0.28–1.01) down to 0.49 (95 per cent confidence interval:0.25–0.94) when deleted case is either of rank

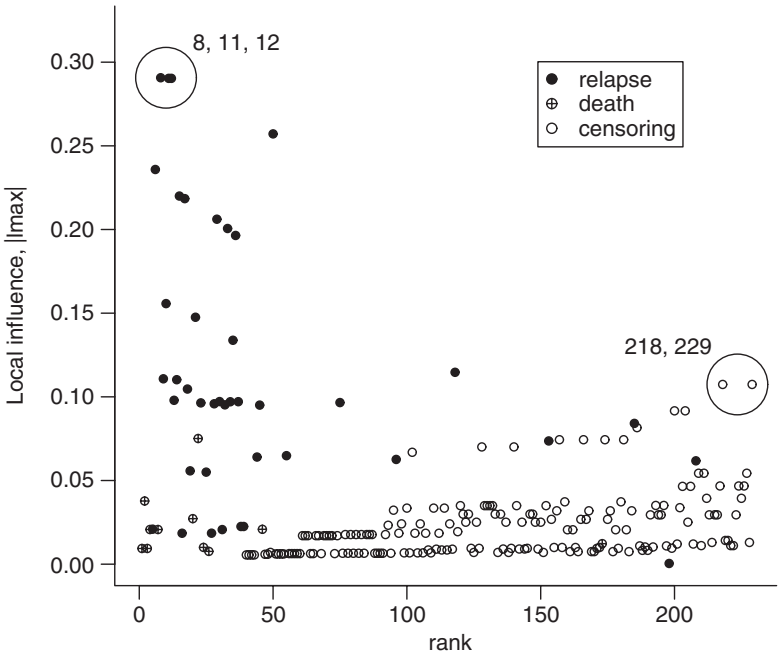


Figure 1. APL93 trial. Plot against failure rank of the local influence ($|l_{\max}|$) in the multivariable Fine and Gray model for the subdistribution hazard of relapse, according to the competing-risks failures, namely relapse and death, and censoring.

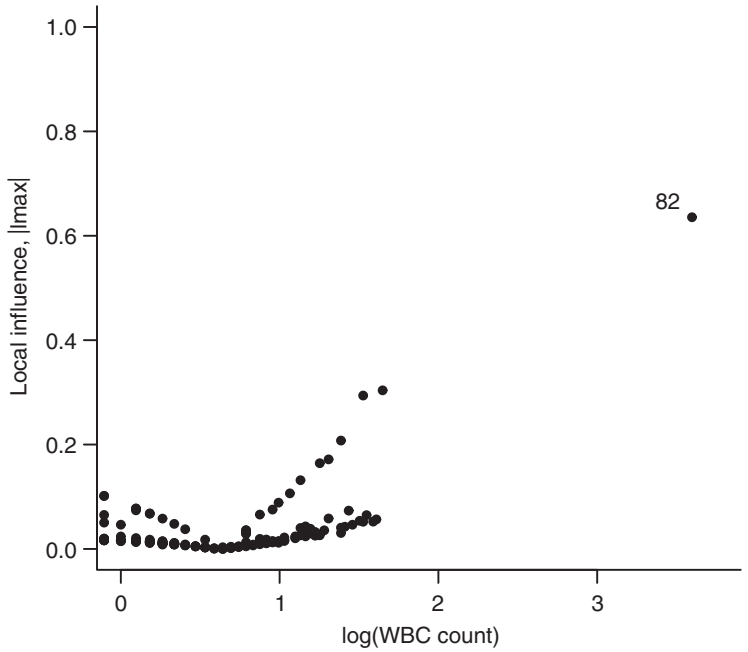


Figure 2. APL93 trial. Plot of local influence ($|l_{\max}|$) against logarithm of WBC count in the univariable Fine and Gray model for the subdistribution hazard of relapse incorporating WBC as a covariate.

8, 11 or 12, and down to 0.52 (95 per cent confidence interval:0.27–0.99) when deleted case is either of rank are 218 or 229, suggesting at least the instability of statistical inference regarding this effect.

Finally, we plotted in Figure 2 the local influences for the univariable Fine and Gray model incorporating baseline white blood cells (WBC) count against WBC values. The local influence of the patient with rank 82 appears rather different from the others. $C_{\max}$ was equal to 0.60 indicating no evidence of local sensitivity. We checked the WBC value of this patient, which turned out to be 42.7 Giga/l, while all other patients had WBC values below 5.2 Giga/l. Therefore, this patient could appear as an outlier with regards to covariate value. This underlines the interest of influential analysis in such a detection.

SIMULATION

In this section, we present the results of numerical investigations to evaluate the importance of assessing local influence as a function of failure ranks, in the setting of two competing risks.

Failure times were generated as described in Reference [10]. Briefly, the subdistributions for cause 1 failures were given by

$$\Pr(T_i \leqslant t, \varepsilon_i = 1|Z_i) = 1 - [1 - p\{1 - \exp(-t)\}]^{\exp(Z_i\beta_1)}$$

which is a unit exponential mixture with mass $1 - p$ at ∞ when $Z = 0$, and uses the proportional subdistribution hazards model to obtain the subdistribution for non-zero covariate values. The subdistribution for cause 2 failures was then obtained by taking $\Pr(\varepsilon_i = 2|Z_i) = 1 - \Pr(\varepsilon_i = 1|Z_i)$ and using an exponential distribution with rate $\exp(Z_i\beta_2)$ for $\Pr(T \leqslant t|\varepsilon_i = 2, Z_i)$. We generated the covariate Z as Bernoulli (0.5) variates. Censoring times were generated from an exponential distribution with rate 0.25. We used the true parameter values of $(\beta_1, \beta_2) = (1, 0)$, and $p = 0.62$, with 20 000 samples of size $n = 80$. This gave 64 per cent cause 1 failures, 18 per cent cause 2 failures, and 18 per cent of censoring. From the samples, we computed the median of local influences for each rank of failure, according to failure cause (or censoring) and covariate value.

In Figure 3, we plotted $|l_{\max}|$ against ranks of observations according to failure cause and covariate values. Local influence appears related to the rank for all causes of failure, although with marked differences. In case of failure from the cause of interest, the influence decreased with the rank and then, re-increased, though with shape depending on Z value. In case of failure from the competing cause, local influence monotonically increases over the rank with slope varying according to the covariate value. Finally, in case of censoring, local influence increases over ranks from null influence up to the maximal local influence of the observations that fail from the competing cause. Actually, these findings could be expected from formula (3) as detailed in the Appendix.

DISCUSSION

We have shown how to extend and use likelihood displacement and local influence to detect influential cases when one is faced with competing-risks censored data. Local influence appears a useful tool for identifying influential cases and for assessing the effects that small

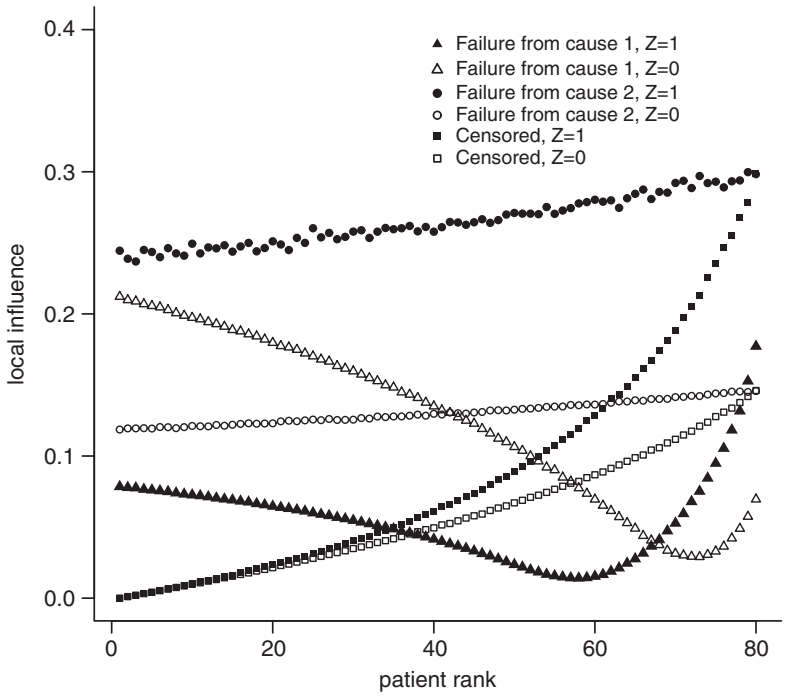


Figure 3. Simulations. Plot against failure rank of the local influence in the univariable Fine and Gray model for the subdistribution hazard of failure from cause 1, according to failure cause and censoring, and covariate value.

perturbations to the assumed data/model will have on inferences. The methods are computationally simple and the results are easy to display graphically. They give useful information about model or data that need extra scrutiny. Actually, our findings exhibited possible outliers with respect to the covariate values and indicated the effect of the rank of failure (or censoring) on the local influence.

The importance of rank in the local influence was suggested by the findings of the real data set, then confirmed on the simulation study. Of note, a similar behaviour of increasing curves over ranks was observed for the Cox model in the absence of competing risks (data not shown). This could be paralleled to the high rank of individual influences reported from such analysis by Cain and Lange [5]. We do think that the impact of such a behaviour on the subsequent inference is of prime interest, though difficult to assess directly. It, at least, appears related to the measures of influence in such models, that are basically summations over risk sets the size of which is heavily dependent on ranks. Further research is required in this setting to consider this issue.

Once outliers and influential observations have been identified, one should suggest the application of the data-analytic strategy. A naive outlier-deletion method, that is re-testing for covariate effects after exclusion of influential cases, is often reported [5, 9], though it could lead to invalid inference [20, 21]. To compensate for the number of comparisons being made, powerful procedures for multiple significance testing could be used [22]. Nevertheless, how to treat such influential data points depends on many considerations [23], so that universal

recommendations have been reported elusive [2]. In any case, influential observations provide information with regards to the statistical model or data. Published reports that do not mention such influence studies could mislead the reader as to the reliability of their conclusions [5]. Finally, when estimating effect size from a clinical trial, one should kept in mind that intent-to-treat analysis is the primary concern of the study, based on robust method when available. These robust methods are designed to be insensitive to selected aspects of the model or data. However, as far as we know, no robust method has been proposed for competing-risks data. Therefore, we devote most of the paper to diagnostics for case-weight influence and we do not consider the misspecification of the model assumptions. Influential analyses should appear as explanatory analyses that could raise the question of instability of the estimated effect size with regard to some minor perturbations of the model or data.

We extended local influence measures to approximate the influence of individual cases upon regression coefficient estimates obtained from the Fine and Gray model. These local influence measures appear dependent on the scale of the perturbation weights. Other influence approaches could have been used. Residual analyses have been proposed in the setting of the Cox proportional hazards model [9]. However, plots involving these residuals are generally difficult to interpret. Finally, they make up a part of the case weighted measures developed in this paper. In 1992, Escobar end Meeker compared the diagonal elements of the matrix $\Delta^{\mathrm{T}} I^{-1} \Delta$ to $\chi^2(1-\alpha, p+2)$ to detect influential observations, but the test appeared very conservative [3]. They insisted on examining all of the eigenvectors and particularly those with large eigenvalues to identify all locally influential cases. Recently, Zhu and Zhang [8] proposed a new promising diagnostic procedure using local influence, with a simple statistic based on the normal curvature for model misspecification. It has been applied to generalized linear models, where the log likelihood function consists of a sum of terms, one for each observation, i.e. where $L(\beta|\omega) = \sum_{i=1}^{n} \omega_i \log p(y_i; \beta)$. However, in the setting of regression models for right-censored and even competing-risks data, the log likelihood is a sum over risk sets and not over all observations. Thus, for these models, the estimation of case influence is somewhat more complex, and this statistic does not appear to apply directly. Further developments for survival models appear promising.

APPENDIX: INTERPRETATION OF THE PLOTS OF LOCAL INFLUENCE AGAINST FAILURE (CENSORING) RANK

Using the expression of the change in score vector induced by minor perturbation ω (equation (3)), the shape of local influence curves ($|l_{\max}|$) over ranks should have been expected from the behaviour of $|\Delta|$ over ranks.

Let the ith element of $|\Delta|$ refer to the change in score used in measuring the local influence of the ith observation. Since observations are ordered according to failure (censoring) times, ith observation is that with the ith rank of failure (censoring): $|\Delta_1|$ denotes the change in score due to the minor perturbation of the observation with the first rank of failure (censoring) while $|\Delta_n|$ denotes that due to the minor perturbation of the observation with the last rank of failure (censoring). The shape of local influence curves, for fixed covariate value, displayed in Figure 3 appears to be related to the cause of failure, as follows:

- For patients who failed from the cause of interest ($\varepsilon_i = 1$), local influence curves exhibited a minimum. This minimum value is related to the use of an absolute value as the

measure of local influence. Suppose, without loss of generality, that the first and the last observations have failed from the cause of interest. Since Δ_1 and Δ_n have opposite signs, the resulting absolute values $|\Delta_1|$ and $|\Delta_n|$ are positive, with minimum value over the ranks.

- For patients who failed from the competing cause ($\varepsilon_i = 2$), local influence curves increased monotonically over ranks. Actually, as $|\Delta_{i,1}|$ is null for those patients who did not experienced the outcome of interest, $|\Delta_i|$ restricts to $|\Delta_{i,2}|$. Moreover, in this case, as mentioned in the main text, the set D_i over which is defined $|\Delta_{i,2}|$ includes all individuals. Therefore, differences in $|\Delta_{i,2}|$ only result from differences in the weights v_{il}. Since $X_i \leqslant X_j$ implies $v_{il} \leqslant v_{jl}$, $|\Delta_i|$ is an increasing function of the rank. Of note, in absence of censoring, $v_{il} = 1$, so that Δ_i is constant over ranks.
- Finally, for censored patients (i.e. when $\varepsilon_i = 0$), a similar monotonic increase of local influence was observed over ranks, reaching the same maximum influence for the last observation than that observed with the last competing failure individual. Indeed, as all patients who did not experience the failure cause of interest, those patients shared a $|\Delta_i|$ expression that is restricted to $|\Delta_{i,2}|$, but with D_i sets increasing over ranks. If the first observation is censored, the set D_1 restricts to this first patient; thus, $|\Delta_{1,2}| = 0$. If the last observation is censored, the set D_n includes all individuals. Thereby, the local influence curves reaches the same influence point that previously observed if the last observation has failed from a competing cause.

REFERENCES

1. Cook RD, Weisberg S. *Residuals and Influence in Regression*. Chapman & Hall: London, 1982.
2. Cook RD. Assessment of local influence. *Journal of the Royal Statistical Society B* 1986; **48**:133–169.
3. Escobar L, Meeker W. Assessing influence in regression analysis with censored data. *Biometrics* 1992; **48**: 507–528.
4. Chatterjee S, Hadi A. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science* 1986; **1**:379–416.
5. Cain KC, Lange NT. Approximate case influence for proportional hazards regression model with censored data. *Biometrics* 1984; **40**:493–499.
6. Reid N, Crépeau H. Influence functions for proportional hazards regression. *Biometrika* 1985; **72**:1–9.
7. Chen CH, Wang PC. Diagnostic plots in Cox's regression model. *Biometrics* 1991; **47**:841–850.
8. Zhu H, Zhang H. A diagnostic procedure based on local influence. *Biometrika* 2004; **91**:579–589.
9. Pettitt AN, Bin Daud I. Case-weighted measures of influence for proportional hazards regression. *Applied Statistics* 1989; **38**:51–67.
10. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 1999; **94**:496–509.
11. Machin D. On the evolution of statistical methods as applied to clinical trials. *Journal of Internal Medicine* 2004; **255**:521–528.
12. Andersen PK, Abildstrøm SZ, Rosthøj S. Competing risks as a multi-state model. *Statistical Methods in Medical Research* 2002; **11**:203–215.
13. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics* 1978; **34**:541–554.
14. Gray RJ. A class of *k*-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics* 1988; **116**:1141–1154.
15. Pepe S. Inference for events with dependent risks in multiple endpoints studies. *Journal of the American Statistical Association* 1991; **86**:770–778.
16. Fine JP. Regression modelling of competing crude failure probabilities. *Biostatistics* 2001; **2**:85–97.
17. Klein J, Andersen P. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* 2005; **61**:223–229.
18. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982; **69**:239–241.

19. Fenaux P, Chastang C, Chevret S, Sanz M, Dombret H, Archimbaud E, Fey M, Rayon C, Huguet F, Sotto JJ, Gardin C, Makhoul PC, Travade P, Solary E, Fegueux N, Bordessoule D, Miguel JS, Link H, Desablens B, Stamatoullas A, Deconinck E, Maloisel F, Castaigne S, Preudhomme C, Degos L. A randomized comparison of all transretinoic acid (ATRA) followed by chemotherapy and ATRA plus chemotherapy and the role of maintenance therapy in newly diagnosed acute promyelocytic leukemia. *Blood* 1999; **94**:1192–1200.
20. Dupuis D, Hamilton D. Regression residuals and test statistics: assessing naive outlier detection. *Canadian Journal of Statistics* 2000; **28**:259–275.
21. Welsh A, Ronchetti E. A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference* 2002; **103**:287–310.
22. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Statistics in Medicine* 1990; **9**:811–818.
23. Belsley A, Kuh E, Welsch R. *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. Wiley: Hoboken, 1980.

3.5 Comparaison des modèles pour risques compétitifs via leur mesure d'influence locale

Dans le cadre des risques compétitifs, nous avons vu que l'approche basée sur un modèle de Cox pour la fonction de risque cause-spécifique et l'approche basée sur le modèle de Fine et Gray pour la fonction de risque de sous-répartition [12] étaient deux approches alternatives.

Pour mieux appréhender les différences de modélisation issues de ces deux stratégies, nous avons étudié l'influence individuelle des observations dans ces deux modèles. Les mesures d'influence locale sont en effet un outil intéressant pour identifier les observations ayant une influence importante et pour évaluer les effets sur l'inférence de modifications des données [5, 49]. Ainsi, l'intérêt principal de ces mesures n'est pas de définir le meilleur modèle, mais de fournir une meilleure compréhension de la structure des données dans les deux modèles de régression [6].

3.5.1 Méthodes

Nous avons étudié l'influence locale de ces deux modèles par une étude de simulation. La méthode de simulation est la même que celle exposée plus haut, en absence de censure.

3.5.2 Résultats

3.5.2.1 Influence individuelle en fonction du rang

La figure 3.3 illustre, pour $\beta = 1$, $n = 200$ (effectif associé à une puissance théorique d'environ 0.8) et $p' = 0.7$, l'influence locale en fonction du rang pour le modèle de Cox et celui de Fine et Gray. On constate que l'influence locale dépend du rang d'observation pour tous les individus dans un modèle de Cox, mais uniquement pour les individus ayant présenté l'événement d'intérêt dans le modèle de Fine et Gray. En effet, pour les individus n'ayant pas présenté l'événement 1, il n'y a pas de modification de l'influence locale selon le rang d'observation pour une même valeur de X dans le modèle de Fine et Gray. Dans un modèle de Cox, ces mêmes observations ont une influence locale qui croît lentement avec le rang de l'observation, mais avec une augmentation très importante et brutale pour les derniers rangs.

Dans les deux modèles, l'influence locale des observations ayant développé un événement 1 décroît dans un premier temps avec le rang. Elle remonte de façon brutale dans le modèle de Cox pour les derniers rangs. Pour le modèle de Fine et Gray, elle continue de décroître, avec une augmentation terminale uniquement pour des prévalences de l'événement 1 élevées (figure 3.4).

D'une manière plus générale, comme pour les individus ayant présenté l'événement 2, le modèle de Cox donne aux derniers individus ayant présenté l'événement 1 une importance beaucoup plus grande que celle accordée dans le modèle de Fine et Gray.

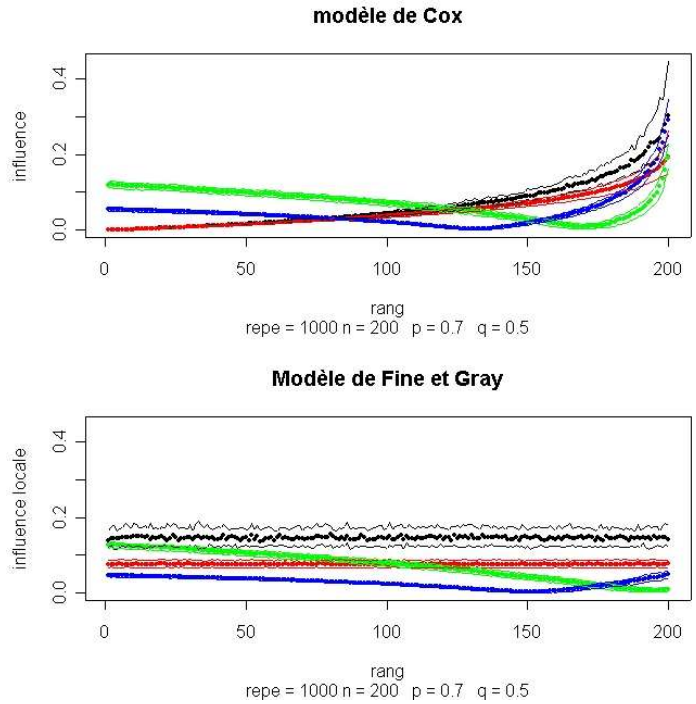
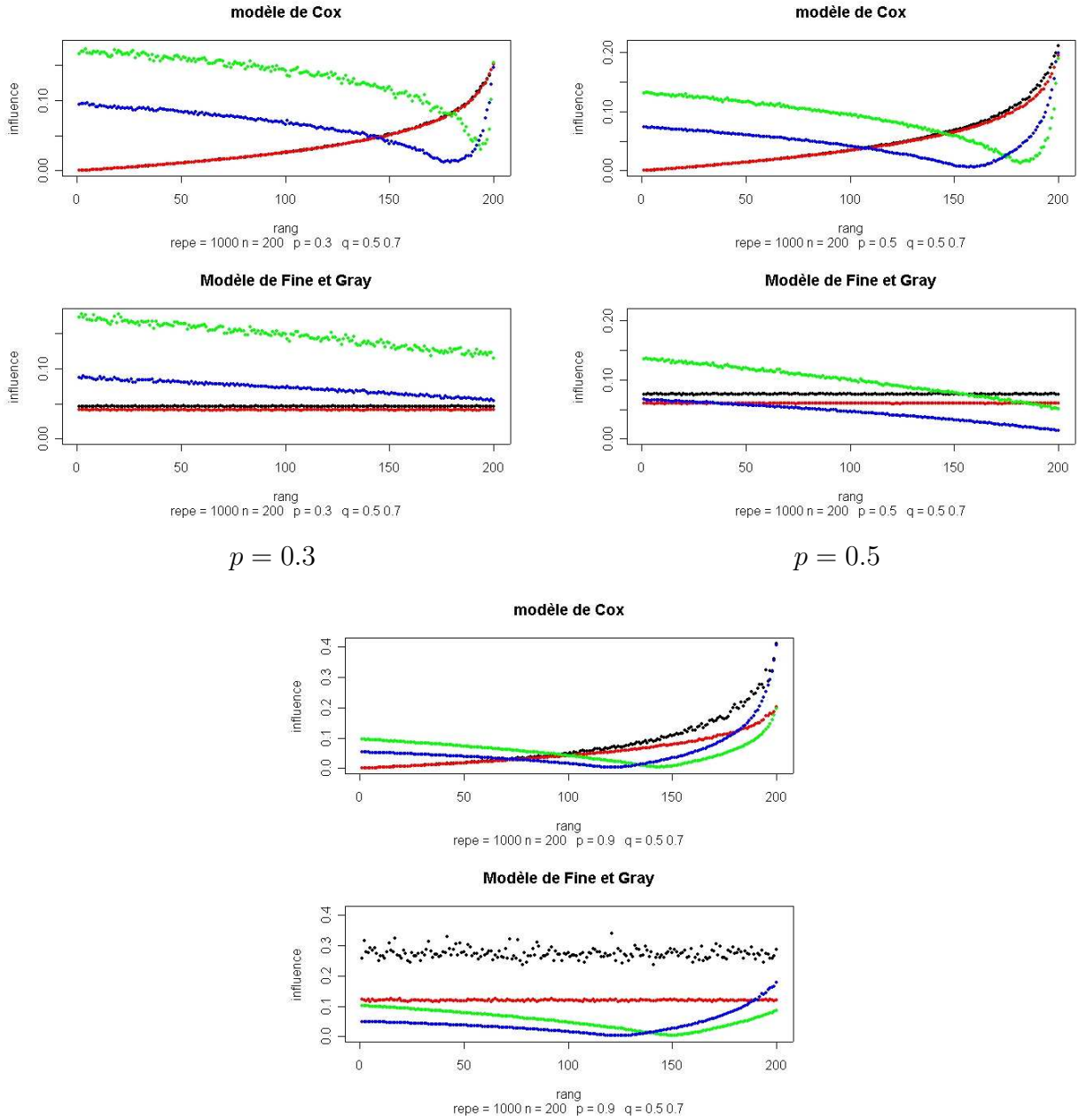


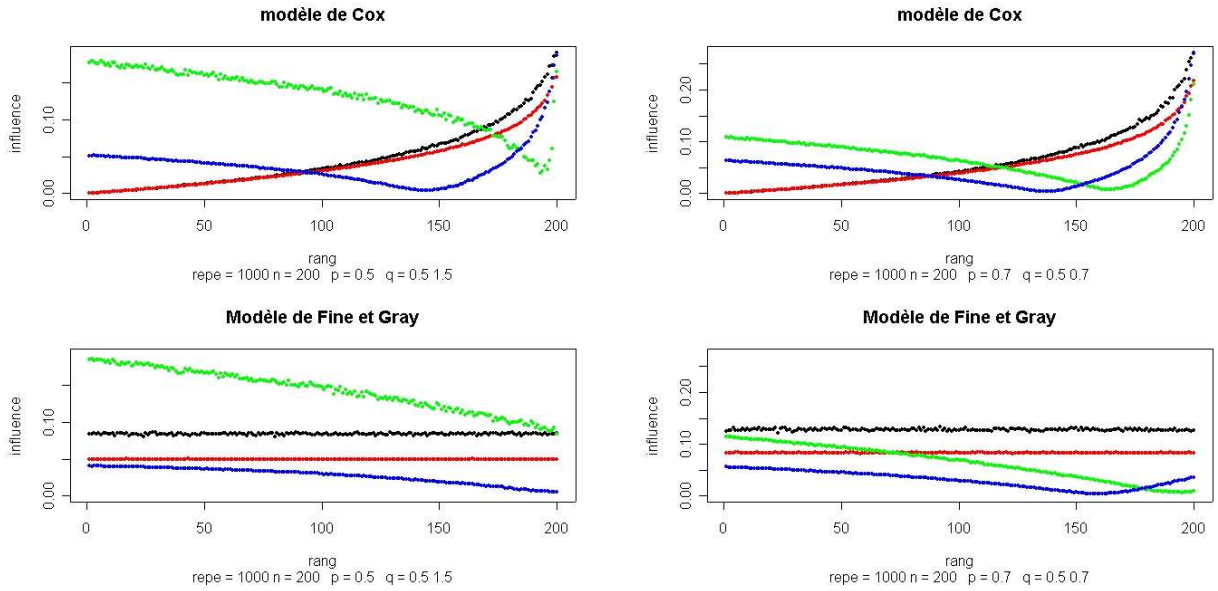
FIG. 3.3 – **Comparaison de l'influence locale des observations en fonction du rang selon le type d'événement ε et la covariable X, dans un modèle de Cox et dans un modèle de Fine et Gray**. $\beta = 1$, $p = 0.7$, $n = 200$, $N = 1000$. $\varepsilon = 1$ et $x = 1$ (•), $\varepsilon = 1$ et $x = 0$ (•), $\varepsilon = 2$ et $x = 1$ (•), $\varepsilon = 2$ et $x = 0$ (•). Les courbes en trait fin correspondent aux 10ème et au 90ème percentiles.



$p = 0.3$ $p = 0.5$

$p = 0.9$

FIG. 3.4 – **Modification de l'influence locale des observations selon le rang, le type d'événement ε et la covariable X, dans un modèle de Cox et dans un modèle de Fine et Gray en fonction de la prévalence de l'événement d'intérêt**. $\beta = 0.7$, $n = 200$, p est respectivement égal à 0.3, 0.5 et 0.9, 1000 répétitions. $\varepsilon = 1$ et $x = 1$ (●), $\varepsilon = 1$ et $x = 0$ (●), $\varepsilon = 2$ et $x = 1$ (●), $\varepsilon = 2$ et $x = 0$ (●).



$\beta = 1.5$ $\beta = 0.7$

FIG. 3.5 – **Comparaison de la médiane de l'influence locale des observations dans un modèle de Cox et dans un modèle de Fine et Gray**. $\beta = 1.5$ à gauche et $\beta = 0.7$ à droite, $p = 0.7$, $n = 200$, 1000 répétitions. $\varepsilon = 1$ et $x = 1$ (●), $\varepsilon = 1$ et $x = 0$ (●), $\varepsilon = 2$ et $x = 1$ (●), $\varepsilon = 2$ et $x = 0$ (●).

L'augmentation de la prévalence de l'un des deux événements conduit à diminuer, quelque soit le modèle, l'influence locale des individus ayant présenté cet événement. Ceci est une conséquence directe du fait que la somme des carrés des influences locales est égale à 1.

Ces résultats ne semblent pas modifiés par la valeur du coefficient de régression β (figure 3.5), ni par la taille de l'échantillon (figure 3.6).

Cependant, pour un échantillon donné, si on tronque systématiquement les dernières observations de l'échantillon, et que l'on ré-estime l'influence locale, on s'aperçoit que des individus ayant une très faible influence sur l'échantillon complet deviennent très influençants après troncature, comme l'illustre la figure 3.7. L'influence d'un individu semble donc dépendante de son rang relatif (rang de l'individu rapporté à la taille de l'échantillon). Pour étudier la position relative du minimum de la fonction d'influence locale selon le rang chez les individus ayant développé l'événement d'intérêt, nous avons réalisé une étude de simulation : sur chaque échantillon généré selon la méthode précédemment décrite, nous avons tiré au sort la taille n selon une loi uniforme discrète sur $[50, 200]$. On a alors calculé le rang relatif de l'observation avec $\varepsilon = 1$ ayant l'influence locale minimale pour chaque valeur de X, et réitéré $N = 1000$ échantillons indépendants. La moyenne du rang relatif de l'observation ayant l'influence locale minimale était de 0.90 (écart-type=0.05) pour $X = 0$, et de 0.78 (écart-type=0.05) pour $X = 1$.

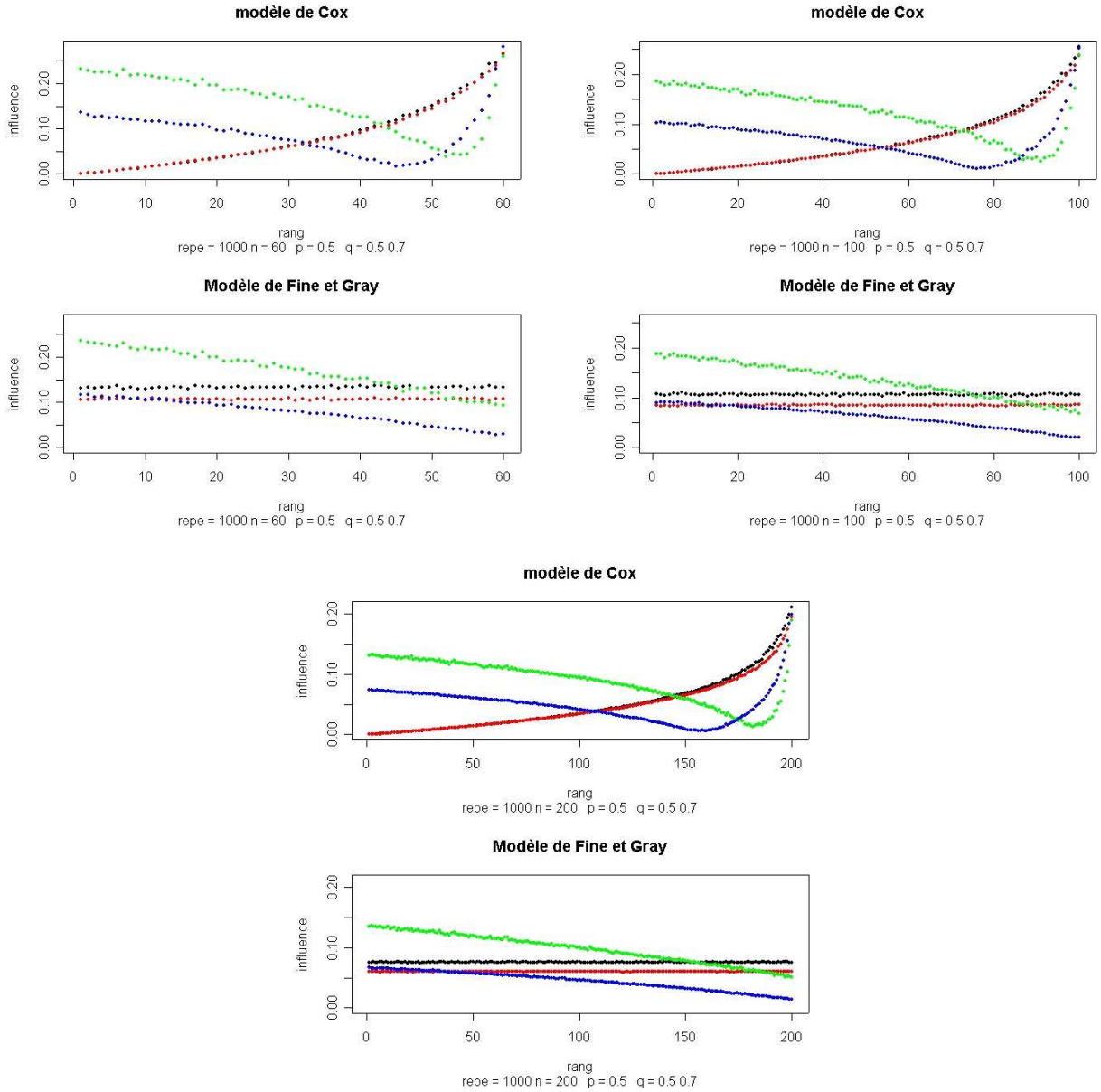


FIG. 3.6 – **Comparaison de la médiane de l'influence locale des observations dans un modèle de Cox et dans un modèle de Fine et Gray**. $\beta = 0.7$, $p = 0.93$, n est respectivement égal à 60, 100 et 2000, 1000 simulations. $\varepsilon = 1$ et $x = 1$ (●), $\varepsilon = 1$ et $x = 0$ (●), $\varepsilon = 2$ et $x = 1$ (●), $\varepsilon = 2$ et $x = 0$ (●).

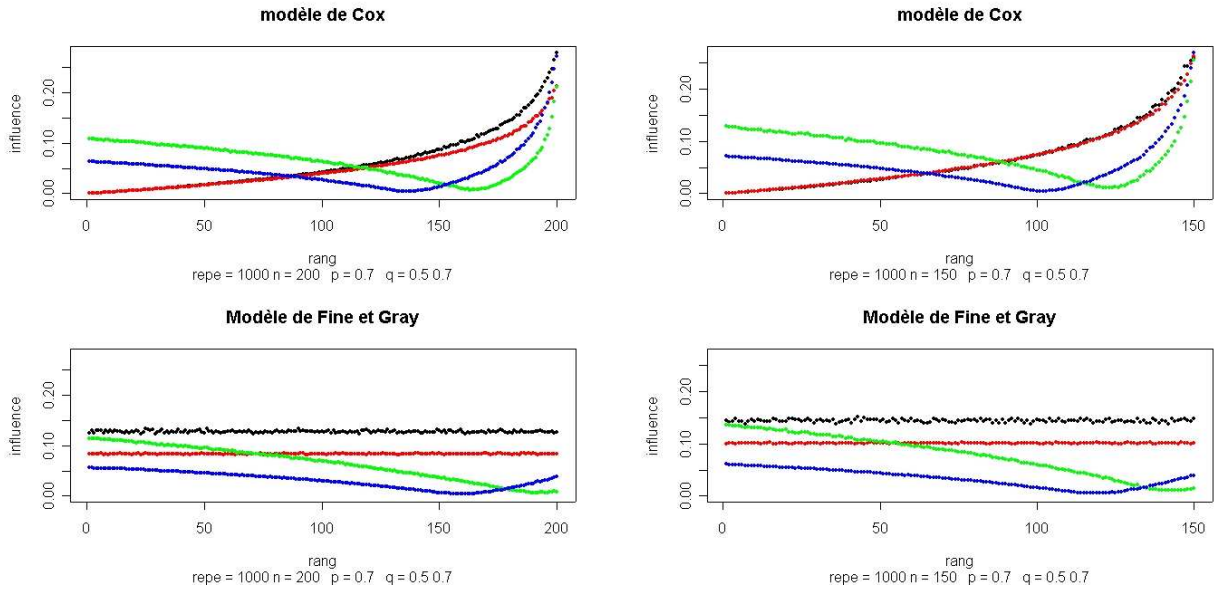


FIG. 3.7 – **Influence locale des observations dans un modèle de Cox et dans un modèle de Fine et Gray, avant ($n = 200$, à gauche) et après troncature des 50 derniers individus ($n = 150$, à droite**. $\beta = 0.7$, $p = 0.7$, $N = 1000$. $\varepsilon = 1$ et $x = 1$ (●), $\varepsilon = 1$ et $x = 0$ (●), $\varepsilon = 2$ et $x = 1$ (●), $\varepsilon = 2$ et $x = 0$ (●).

3.5.2.2 Implications

Nous nous sommes intéressés alors aux implications, en termes de conclusion du test (H_0 : $\beta = 0$), des observations de rang élevé. Le mode de simulation est identique au précédent, mais nous nous sommes focalisés sur les modifications que pouvait entraîner le retrait systématique du dernier individu de l'échantillon sur les conclusions du test. Le tableau 3.1 indique pour trois valeurs de β (=0.7,1,1.5) et pour une puissance théorique de 0.8 à 0.95 [10, 51] le nombre de fois où sur 1000 simulations la conclusion est modifiée par le retrait du dernier sujet (en termes de rejet ou de non rejet de H_0). Il apparaît clairement dans ce tableau que pour une puissance théorique de l'ordre de 0.8, la conclusion avec le modèle de Cox est sensible au retrait du dernier sujet avec un pourcentage de conclusions modifiées proche de 10%, supérieur à celui observé pour le modèle de Fine et Gray (3%).

Modèle de Cox									
	β **théorique** 0.7			β **théorique** 1			β **théorique** 1.5		
puissance théorique	*0.8*	*0.9*	*0.95*	*0.8*	*0.9*	*0.95*	*0.8*	*0.9*	*0.95*
nbre de sujets nécessaire	193	258	319	93	124	154	41	54	67
puissance estimée	0.78	0.88	0.94	0.79	0.88	0.93	0.76	0.85	0.93
puissance estimée sans le dernier individu	0.78	0.94	0.93	0.78	0.88	0.91	0.73	0.83	0.92
beta médian estimé β^C	0.57	0.56	0.55	0.82	0.80	0.82	1.28	1.27	1.27
beta médian estimé sans le dernier individu $\beta^C_{(n)}$	0.57	0.57	0.55	0.83	0.82	0.80	1.26	1.27	1.25
probabilité de modifier les conclusions sans le dernier individu	0.08	0.03	0.02	0.12	0.06	0.04	0.13	0.10	0.05
Modèle de Fine et Gray									
	β **théorique** 0.7			β **théorique** 1			β **théorique** 1.5		
puissance théorique	*0.8*	*0.9*	*0.95*	*0.8*	*0.9*	*0.95*	*0.8*	*0.9*	*0.95*
nbre de sujets nécessaire	128	171	212	63	84	104	28	38	47
puissance estimée	0.80	0.89	0.96	0.79	0.88	0.94	0.75	0.87	0.94
puissance estimée sans le dernier individu	0.79	0.88	0.96	0.79	0.88	0.94	0.73	0.86	0.93
beta médian estimé β^F	0.71	0.70	0.70	1.03	1.02	1.00	1.50	1.51	1.52
beta médian estimé sans le dernier individu $\beta^F_{(n)}$	0.71	0.69	0.70	1.02	1.01	1.00	1.50	1.48	1.50
probabilité de modifier les conclusions sans le dernier individu	0.03	0.02	0.01	0.03	0.02	0.01	0.07	0.04	0.02

TAB. 3.1 – Pourcentage de conclusions modifiées par le retrait systématique du dernier individu de l'échantillon, dans les modèles de Cox et de Fine et Gray. $N = 1000$.

3.6 Conclusions

L'objectif de ce travail était d'étudier deux modèles décrivant le risque de survenue d'un événement d'intérêt dans un contexte de risques compétitifs. Plutôt que de se placer dans une stratégie de sélection de modèle, notre idée a été d'utiliser les mesures d'influence individuelle pour mieux comprendre les implications du choix d'un des deux modèles en termes d'influence de chaque observation, qu'elle ait présenté l'événement d'intérêt ou non [6]. En absence de mesure d'influence locale pour le modèle de Fine et Gray, notre premier travail a consisté à étendre une mesure décrite pour le modèle de Cox au modèle de Fine et Gray. Cette extension a nécessité la définition de l'ensemble D_i pour une situation de compétition.

Une étude de simulation nous a conduit à deux résultats principaux quant à l'influence différente des individus dans les deux modèles.

Le premier résultat concerne, dans le modèle de Cox, le rôle joué par le rang de l'observation, à valeur fixe de la covariable. Ainsi, l'influence individuelle est maximale pour les observations des derniers rangs, qu'elles aient ou non présenté l'événement d'intérêt. De plus, à X fixe, les dernières observations censurées semblent d'influence supérieure à celles des observations non censurées, et à événement fixe (pour $\beta \geq 0$), les dernières observations avec $X = 1$ semblent avoir une influence supérieure aux dernières observations avec $X = 0$. Ce résultat semble corroboré par les données de la littérature. Ainsi, sur un exemple décrit par Pettit et Bin Daud ([23], figure 2), les trois observations rapportées comme ayant les influences locales les plus grandes sont les trois dernières observations non censurées de l'échantillon, toutes avec $X = 1$. De même, sur l'exemple présenté dans l'article de Cain et Lange ([9], figure 1C), les trois individus d'influence locale maximale sont les trois dernières observations censurées de l'échantillon, toutes également avec $X = 1$. Enfin, Nardi [52] a observé que les individus « vivant beaucoup trop longtemps » étaient particulièrement influençants, en utilisant une mesure d'influence basée sur des résidus. L'aspect structurel, modèle-dépendant, du rôle du rang dans l'influence locale a été illustré dans notre travail par la position stable, à X fixe, du minimum de la fonction d'influence locale pour $\varepsilon = 1$ simulée (ce minimum correspondant à un changement de signe du $lmax$). Cependant, aucun des auteurs ne semble avoir considéré le rôle structurel du rang dans ces influences élevées. Les diverses définitions ou interprétations des observations ayant une influence individuelle importante (ou *outliers*) pourraient cependant intervenir dans ce constat, puisque ce terme regroupe des sujets avec données extrêmes voire aberrantes (ou « contaminants »), que ce soit du fait d'un mélange de distributions ou de données erronées, et des observations reflétant le choix d'un modèle inadapté ou mal spécifié ([53], chapitre *outlier*). Dans la mesure où notre étude de simulation a été basée sur un modèle à risques compétitifs, on pourrait aussi se demander si le rôle du rang dans l'influence individuelle observée dans le modèle de Cox n'est pas uniquement le reflet d'un modèle inadapté. Cependant, même en simulant des observations à partir d'un modèle

de Cox, nous avons observé des résultats similaires. Enfin, la plupart des auteurs ont considéré un modèle avec des covariables continues, pour lesquelles les diverses sources d'influence peuvent s'additionner. Ainsi, Barlow [46] a rapporté que les mesures d'influence globale permettaient d'identifier des observations influençantes soit du fait de valeurs extrêmes pour les covariables, soit du fait d'une durée d'exposition très courte ou très longue.

Le second résultat concerne les différences d'influence locale observées dans un modèle de Fine et Gray par rapport au modèle de Cox, avec d'une part l'absence de rôle du rang des observations n'ayant pas présenté l'événement d'intérêt, et d'autre part, une influence individuelle des observations ayant présenté l'événement d'intérêt proche de celle des mêmes individus pour le modèle de Cox, sauf pour les derniers rangs. L'absence de rôle sur l'influence individuelle du rang des observations n'ayant pas présenté l'événement d'intérêt était attendue, compte-tenu de des définitions respectives des ensembles R_i et D_i. Enfin, pour les observations ayant développé l'événement d'intérêt, les mesures d'influence locale sont proches pour les deux modèles, excepté dans les rangs élévés où l'influence dans le modèle de Cox est supérieure à celle de la même observation dans le modèle de Fine et Gray. Cependant, lorsque la prévalence de l'événément d'intérêt tend vers 1, l'influence de ces derniers événements dans le modèle de Fine et Gray se rapproche de celle observée dans le modèle de Cox, soulignant la proximité des deux modèles dans cette situation où, de fait, il n'existe pratiquement plus de compétition.
Dans les deux modèles, l'impact de l'omission de ces sujets influençants sur les conclusions du test de l'effet de la covariable à zéro n'est enfin pas négligeable, comme l'a illustré le tableau 3.1.

Au delà du débat sur le choix de l'approche à choisir pour analyser des données de survie en présence de compétition, ce travail nous a donc permis de souligner que de choisir une approche cause-spécifique conduit à donner une influence plus grande aux individus ayant les temps d'exposition les plus longs de l'échantillon. Au contraire, privilégier la fonction de sous-répartition conduit à attribuer une influence plus stable à toutes les observations (à X fixe).

Chapitre 4

Influence individuelle et essais séquentiels

4.1 Essais séquentiels de recherche de dose

4.1.1 Contexte

Les essais de phases précoces, essais de phase I et/ou II, sont les premiers essais cliniques réalisés chez les êtres humains, en règle des sujets volontaires sains, par opposition aux études pré-cliniques réalisées *in vitro* ou chez l'animal. L'objectif majeur de ces essais, encore appelés essais de recherche de dose, est de définir une dose médicamenteuse sur des critères de tolérance et/ou d'efficacité permettant la poursuite des investigations thérapeutiques par des essais de phase III. En pratique, ils estiment, sur des échantillons de petite taille, en fonction de la dose, la probabilité de survenue d'une réponse en tout ou rien mesurée en règle en termes de toxicité ou de tolérance, et parfois d'efficacité thérapeutique [54].

Les essais de phase I sont centrés sur l'étude de la relation dose-toxicité du nouveau médicament. Ils sont conduits chez le volontaire sain, avec pour objectif principal l'étude de la tolérance du médicament, sur des critères en règle pharmacocinétiques ou pharmacodynamiques. En hémato-cancérologie, pour des contraintes éthiques évidentes, les essais de phase I sont conduits chez le sujet malade. La toxicité est alors considérée comme le pré-requis à l'efficacité. Du fait de l'hypothèse sous-jacente d'une relation dose-toxicité monotone croissante, l'objectif est de choisir une dose dont la toxicité est suffisamment élevée pour espérer une efficacité et suffisamment faible pour pouvoir être évaluée de façon plus large dans un essai de phase II. Cette dose recherchée est dite "dose maximale tolérée" (DMT) (*Maximum Tolerated Dose, MTD*). La détermination de la DMT doit se faire en exposant le minimum de patients à des doses trop toxiques voire, plus récemment, en exposant le plus de patients possible aux doses efficaces [54].

4.1.2 Schémas de conduite et d'analyse

Comme l'ont souligné Gatsonis et Greenhouse [55], petites tailles des échantillons et considérations éthiques sont les caractéristiques les plus contraignantes des essais de recherche de dose. Elles expliquent le développement de schémas séquentiels et adaptatifs pour ces essais, selon un double principe : (i) inclure les patients séquentiellement dans l'essai en les traitant à des doses croissantes, et (ii) adapter pour chaque nouvelle inclusion la dose à administrer aux informations disponibles. A partir des observations (doses et toxicités éventuelles) réalisées, une règle d'administration des doses est établie.

Ces schémas se sont développés essentiellement pour les essais de phase I en cancérologie. Le plus ancien et le plus simple, toujours très utilisé, est connu sous le nom de schéma standard *'3+3'* [56, 57]. Les français l'appellent volontiers (à tort) schéma de Fibonacci. Cette méthode est non-paramétrique, car elle ne fait aucune hypothèse sous la forme de la relation sous-jacente entre dose et toxicité (en dehors d'une relation monotone croissante). Les doses sont administrées par cohorte de 3 patients et l'essai débute par la dose la plus faible. La règle d'escalade des doses est empirique : en absence de toxicité observée au sein de la cohorte, la dose est augmentée au niveau immédiatement supérieur pour la cohorte suivante ; si une toxicité est observée au sein de la cohorte, une cohorte supplémentaire de 3 patients reçoit la même dose et on évalue l'ensemble des toxicités observées sur les 6 patients inclus : si au moins deux toxicités sont observées, l'essai est arrêté. La DMT est alors définie comme la dernière dose expérimentée, voire parfois la dose immédiatement inférieure (qui prend parfois le nom de "dose recommandée pour les essais de phase II"). On a montré que la DMT ainsi définie avait une probabilité associée de toxicité comprise entre 10 et 29%, avec une moyenne de 20% [56].

Si cette approche et ses multiples variations (schémas 'A+B', schémas *up and down* [58]) ont connu et connaissent encore un certain succès, elles ont été peu à peu abandonnées au profit de méthodes de modélisation paramétrique, c'est à dire décrivant la relation dose-toxicité par un modèle mathématique connu à un (ou plusieurs) paramètre(s) près. Ces méthodes ont, en effet, rapidement montré qu'elles étaient plus efficientes [13, 59], nécessitaient moins de sujets et permettaient une plus grande souplesse dans le choix de la probabilité de toxicité ciblée. Elles définissent alors spécifiquement la DMT comme un percentile de la relation dose-toxicité. De l'inférence sur le(s) paramètre(s) du modèle, qu'elle soit bayesienne [13, 60, 61] ou par maximum de vraisemblance [62]), on définit une règle d'administration des doses basée sur l'optimisation d'une fonction dépendant des estimations et de la probabilité visée (typiquement la distance entre les probabilités de toxicité estimées et la probabilité de toxicité ciblée). Parmi ces méthodes, la première proposée, la plus utilisée (et la plus simple à implémenter) est la méthode de réévaluation séquentielle (MRS) proposée dans sa version initiale par John O'Quigley en 1990 (ou *Continual Reassessment Method, CRM*) [13]. La relation dose-toxicité ne dépend que d'un seul paramètre et

la règle de choix de la dose suivante se fait par minimisation de la distance à la probabilité cible. En 1995, Whitehead proposa une méthode séquentielle bayesienne où la relation dose-toxicité est décrite par un modèle logistique à 2 paramètres et l'optimisation se fait via la variance *a posteriori* des paramètres [60].

4.1.3 Méthode de Réévaluation Séquentielle (MRS) ou Continual Reassessment Method (CRM)

L'objectif de la Méthode de Réévaluation Séquentielle ou *Continual Reassessment Method* [13] est de déterminer la dose maximale tolérée (DMT), définie comme le $(100 \times \pi)^e$ percentile de la relation dose-toxicité. Son principe est de modéliser de façon paramétrique la relation dose-toxicité, de réévaluer après chaque sujet ou cohorte le(s) paramètre(s) du modèle, de déterminer (parmi un choix prédéterminé de k doses) la dose ayant la probabilité de toxicité estimée la plus proche de la cible (π) et de l'administrer à la cohorte suivante (Figure 4.1). L'essai s'arrête après un nombre prédéterminé d'inclusions.

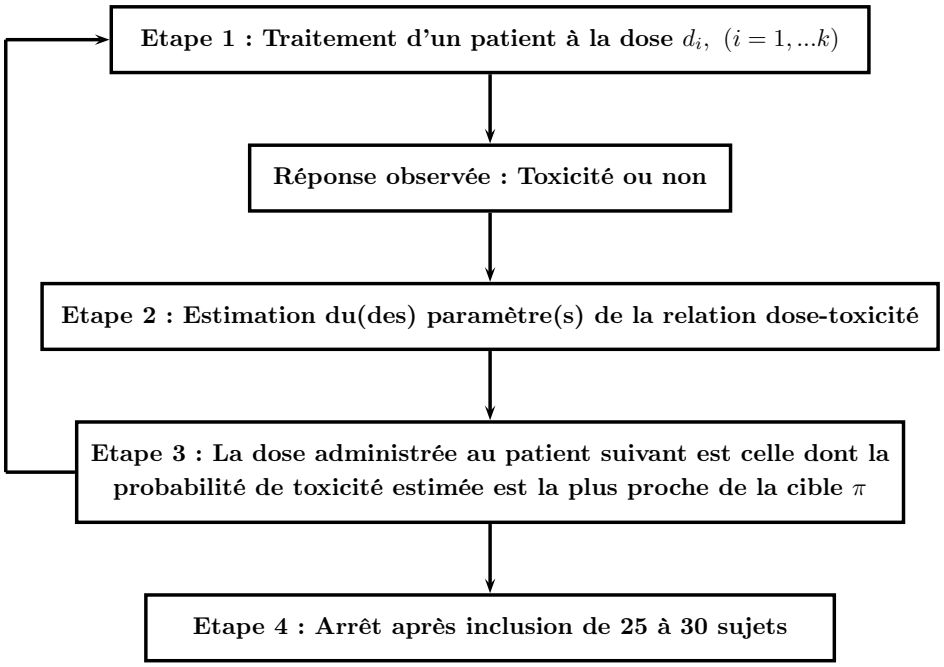


FIG. 4.1 – Représentation schématique de la MRS

4.1.3.1 Formulation et modélisation de la relation dose-toxicité

Soit $\Omega_j = \{x(i), y_i; i = 1, \ldots, j\}$, l'ensemble des observations recueillies après j patients, où $x(i)$ est la dose administrée parmi k doses au $i^{ème}$ patient, and y_i sa réponse (1=toxicité dose-limitante TDL, 0=non TDL).

Soit $\psi(x_i, \theta) = P(Y = 1|x_i)$, $i = 1, ..., k$ une relation dose-toxicité, $\theta \in [0; \infty)$ l'unique paramètre à estimer, et $x_i = \psi^{-1}(p_i, \theta_0)$, où p_i est la probabilité de toxicité initialement supposée associée à la dose réelle d_i $(i = i, \ldots k)$ et θ_0 la valeur initiale de θ.

Si initialement le modèle proposé par O'Quigley [13] était un modèle tangente hyperbolique $\psi(x_i, \theta) = \{(\tanh(x_i) + 1)/2\}^\theta$, $\theta \geq 0$, deux modèles sont aussi très utilisés :

– le modèle logistique [63, 64, 65, 66]

$$\psi(x_i, \theta) = \frac{\exp(a_0 + \theta x_i)}{1 + \exp(a_0 + \theta x_i)},\ \theta \geq 0 \tag{4.1}$$

avec a_0 fixé, le plus souvent $a_0 = 3$ [67].

– le modèle puissance [62]

$$\psi(x_i, \theta) = x_i^\theta,\ \theta \geq 0 \tag{4.2}$$

4.1.3.2 Inférence

L'étape 2 (Figure 4.1) consiste à actualiser, au vu des observations précédentes, la valeur du paramètre θ. Après l'inclusion du $j^{ème}$ patient, la vraisemblance du modèle s'écrit :

$$L_j(\theta) = \prod_{l=1}^{j} \psi(x(l), \theta)^{y_l} (1 - \psi(x(l), \theta))^{(1-y_l)} \tag{4.3}$$

Deux types d'inférence ont été proposés pour l'estimation $\hat{\theta}_{|\Omega_j}$ de θ après j observations.

L'inférence bayesienne, la première décrite [13], considère θ comme une variable aléatoire dont la densité de probabilité *a priori* est notée $g(\theta)$. A noter que θ_0 est alors l'espérance *a priori* de θ. L'espérance *a posteriori* de θ estime alors séquentiellement le paramètre du modèle, ce qui correspond à l'estimateur bayesien de θ pour une fonction de perte quadratique. La densité *a posteriori* de θ après l'observation j s'écrit :

$$f(\theta|\Omega_j) = \frac{L_j(\theta) g(\theta)}{\int_0^\infty L_j(u) g(u) du} \tag{4.4}$$

On obtient donc :

$$\hat{\theta}_{|\Omega_j} = E(\theta|\Omega_j) = \int_0^\infty u f(u|\Omega_j) du = \frac{\int_0^\infty u L_j(u) g(u) du}{\int_0^\infty L_j(u) g(u) du} \tag{4.5}$$

Plusieurs lois *a priori* ont été proposées, les auteurs s'accordant sur l'utilisation d'une loi peu informative, en général la loi exponentielle de paramètre 1 [13, 67].

Une version non bayesienne de la MRS a été secondairement proposée, basée sur l'estimateur du maximum de vraisemblance de θ [62] :

$$\hat{\theta}_{|\Omega_j} = \max_{\theta \geq 0} \{L_j(\theta)\} \tag{4.6}$$

Cet estimateur nécessite l'existence d'une hétérogénéité dans les réponses : il ne peut donc être utilisé avant l'observation de cette hétérogénéité, avant laquelle une règle '3+3' est utilisée.

4.1.3.3 Choix de la dose

Lors de l'étape 3 définissant la règle d'administration des doses (Figure 4.1), la MRS initiale [13] proposait d'attribuer à la première cohorte la dose ayant la toxicité supposée initialement la plus proche de π et d'administrer à la cohorte suivante celle qui minimisait $|\psi(x_i, \hat{\theta}_{|\Omega_j}) - \pi|$. Après j observations, la dose suivante $x(j+1)$ se définit donc par :

$$x(j+1) = \min_{i \in 1,\ldots,k} \left\{ x_i : |\psi(x_i, \hat{\theta}_{|\Omega_j}) - \pi| \right\} \tag{4.7}$$

Plusieurs modifications de cette règle ont été proposées ultérieurement, en règle pour des raisons éthiques ou d'acceptabilité auprès des cliniciens concernés :

- Commencer systématiquement les inclusions par la dose la plus faible [68, 65, 64, 63, 69, 70, 71].
- Inclure les sujets de façon groupée, le plus souvent par cohorte de 3 [68, 65, 70, 71].
- Ne jamais expérimenter une dose supérieure à celle immédiatement au-dessus de la dose la plus haute expérimentée (*no skipping*) [68, 65, 64, 63, 69, 70]

4.1.3.4 Arrêt de l'esssai

Dans les premières versions de la MRS [13], le nombre de sujets à inclure était fixé à l'avance (généralement entre 20 et 30), l'étude s'arrêtant après l'inclusion de ce nombre fixe de sujets (Etape 4, Figure 4.1). De façon à minimiser le nombre de sujets inclus tout en préservant la qualité de l'information obtenue, des règles d'arrêt ont rapidement été proposées :

- Arrêt quand un nombre maximal de 6 patients ont été traités à la DMT estimée [64].
- Arrêt si la probabilité que la dose administrée aux patients suivants reste inchangée est égale à 1 [72].
- Arrêt quand la largeur de l'intervalle de crédibilité de θ se situe dans un intervalle pré-défini et fixe [70].
- Arrêt quand aucun gain n'est attendu dans la précision de l'intervalle de crédibilité de θ [73].

4.1.4 Extension aux essais de Phase II

En présence de contraintes éthiques fortes, le recours à un schéma classique de comparaison de plusieurs doses à l'aide d'un essai randomisé en groupes parallèles expose un nombre important de patients à des doses inefficaces. L'utilisation d'un schéma séquentiel et particulièrement la MRS permet de déterminer une dose minimale efficace (DME) en traitant un nombre plus important de patients à la "bonne" dose et en limitant le nombre de patients ne recevant pas une dose efficace. La MRS s'adapte alors simplement aux essais de Phase II en modélisant la probabilité d'échec (et non pas la probabilité de succès) et en inversant la relation d'ordre des doses [54].

Plusieurs essais de phase II ont ainsi été conduits en pédiatrie et spécialement en néonatologie où les contraintes sur l'efficacité des doses administrées sont encore plus fortes [74, 75, 76, 77, 78]. Les probabilités ciblées sont alors plutôt de l'ordre de 5 à 15% que de 30%.

4.2 Exemple initiateur

Il s'agit d'un essai non publié planifié en 2002 dans le service de biostatistique et informatique médicale de l'hôpital Saint-Louis en collaboration avec une équipe de pédiatres de l'hôpital Saint Vincent de Paul. Il s'agissait d'un essai de phase II ayant pour objectif de déterminer la dose minimale efficace (DME) du Rocuronium, curare non dépolarisant utilisé lors de l'induction anesthésique de manière à faciliter l'intubation des enfants. Les médecins recherchaient une efficacité chez au moins 90% des enfants. Le choix de la MRS a été dicté par le contexte pédiatrique et par le fait que l'innocuité du Rocuronium avait été établie lors d'essais chez l'adulte et d'essais de phase I antérieurs. L'étude fut construite de la façon suivante :

- Probabilité d'échec visée : 0.1, ce qui correspond à une probabilité de succès de la DME de 0.9%
- 6 doses étudiées $d_1 = 1.4,\ d_2 = 1.2,\ d_3 = 1.0,\ d_4 = 0.8,\ d_5 = 0.6,\ d_6 = 0.4\ mg.kg^{-1}$ avec les probabilités d'échec initiales asociées suivantes : 0.001, 0.05, 0.1, 0.3, 0.6, 0.7
- Nombre de patients à inclure : 25
- Cohortes de taille 1
- Choix du modèle logistique (Equation 4.1) pour la paramétrisation de la relation dose-toxicité avec $a_0 = 3$
- L'estimation du paramètre était basée sur une inférence bayesienne avec une densité *a priori* exponentielle de paramètre 1 pour θ

La première dose administrée, selon le schéma initial de la MRS [13] a été la dose d_3, correspondant à une probabilité d'échec initiale supposée de 0.1. Une réponse peu attendue du premier enfant (qui présenta un échec) eut cependant des conséquences sur le déroulement de l'essai. En effet, la MRS recommanda l'administration de la dose la plus élevée (d_1) de Rocuronium et, mal-

gré l'observation exclusive de succès, la recommanda jusqu'au 11ème enfant. L'essai s'arrêta alors pour des raisons indépendantes des résultats. On calcula alors qu'il aurait fallu inclure 4 enfants supplémentaires à la dose maximale (d_1) sans observer d'échec pour pouvoir recommander une dose moindre (d_2).

Le choix d'une inférence bayesienne souleva la question de l'influence de la densité *a priori* choisie. Nous avons donc conduit une analyse rétrospective des données en utilisant l'estimateur du maximum de vraisemblance [62]. Les résultats auraient été proches, avec un total de 21 patients consécutifs traités avec succès à la dose d_1 avant d'expérimenter la dose d_2.

L'influence de la réponse du premier enfant sur le déroulement de l'essai, entraînant l'administration quasi exclusive de d_1 aux 15 premiers patients malgré les succès observés, nous sembla une hypothèse préoccupante à étudier.

4.3 Influence et MRS

Faisant suite à notre travail antérieur sur l'influence individuelle dans le cadre de modèles de survie en présence de compétition, nous avons étudié l'influence des premières observations dans la MRS. Notre objectif était de chercher à mieux comprendre le modèle en identifiant pourquoi cette influence des premières observations semblait si grande, et enfin à proposer une solution visant à diminuer leur impact éventuel.

4.3.1 Influence individuelle dans les essais séquentiels

Certaines caractéristiques de la MRS, méthode séquentielle et adaptative, suggèrent une certaine sensibilité de la méthode à des observations inattendues survenant en début d'essai.

L'objectif de la MRS est de rechercher la dose correspondant à un certain percentile de la relation dose-toxicité. Plus le percentile d'intérêt est faible, plus l'information sur la réponse repose sur un petit nombre d'individus. Ainsi pour les modèles de régression binaire, plus la probabilité de réponse s'éloigne de 0.5, plus les estimations des paramètres du modèle sont sensibles à un petit nombre d'individus [79]. Dans le cas de la MRS, et particulièrement dans le cadre d'un essai de phase II, la probabilité ciblée s'approche de 0.1, donc relativement loin de 0.5. Les effectifs étant faibles, même dans la situation "idéale" où les 30 patients d'un essai reçoivent la dose recherchée, si la probabilité visée est de l'ordre de 0.1, on ne s'attend donc à observer en moyenne que 3 échecs.

Certains auteurs ont déjà souligné la difficulté des méthodes séquentielles à déterminer les quantiles extrêmes. Ainsi Wetherill [80, 81] et plus récemment Roshan [82] ont montré par simulation les faibles performances de la méthode de Robbins-Monro pour déterminer des quantiles

en-dessous de 0.1 ou au-dessus de 0.9. La raison principale avancée pour expliquer ce constat est l'inégalité des mouvements possibles vers le haut et vers le bas [80]. Cet aspect ne peut qu'être renforcé dans le cas de la MRS. En effet, contrairement à la méthode de Robbins-Monro, la MRS ne peut expérimenter qu'un nombre fini de doses. Dans un certain nombre de cas, il devient impossible de descendre "plus bas" ou monter "plus haut". De plus, les doses sont utilisées sur une échelle discrète, introduisant un manque de mobilité supplémentaire.

Enfin, le caractère séquentiel de la MRS implique que l'information apportée par un patient pour ajuster le modèle dose-toxicité est utilisée un nombre de fois inversement proportionnel à son rang d'inclusion. Un patient très influençant au début de l'essai a des risques de l'être sur toutes les estimations qui suivent, renforçant son impact final.

Nous nous sommes donc attachés à montrer l'influence importante des premières observations sur les résultats obtenus avec la MRS. L'objectif était là encore de progresser dans la compréhension de la méthode et de démontrer sa sensibilité à des résultats inattendus précoces.

A notre connaissance, aucun travail ne rapporte de mesure d'influence individuelle appliquée à la MRS. La raison en est sans doute très simple : les mesures existantes d'influence individuelle s'appliquent mal à la MRS. En effet, les mesures les plus simples basées sur des approches *delete-one* où le modèle est ré-estimé en retirant tour à tour toutes les observations ne peuvent s'appliquer à la MRS :

- Retirer un individu revient à éventuellement modifier les doses suivantes à administrer, et donc tout le déroulement ultérieur de l'essai. Il devient alors impossible d'avoir accès à ce nouveau déroulement "virtuel" en dehors des situations de simulation.
- Appliquer simplement à la fin de l'essai les mesures d'influence individuelle ne présente d'intérêt que si l'on souhaite comprendre ce qui s'est passé à la dernière itération, et en aucun cas avant.
- Appliquer les mesures d'influence individuelles à chaque estimation ne répond que de façon brouillonne à notre désir de comprendre le déroulement de l'essai en présence d'observation(s) inattendue(s). Cela conduit à obtenir une quantité importante, difficilement exploitable, de mesures sans pour autant assurer une compréhension de la séquence suivie lors de l'administration des doses.

Nous nous sommes proposés, donc, de montrer par simulation la sensibilité importante de la MRS dans certaines conditions en imposant les réponses de certains individus (et notamment celles des premiers). La méthode semblant *a priori* plus sensible lorsque la probabilité visée est faible, nous nous sommes délibérément placés dans le cadre des essais de phase II.

4.3.2 Etude de simulation

Nous avons simulé selon plusieurs scénarios la réalisation d'un essai de phase II selon la MRS. Les deux types d'inférence (bayesienne et non) ainsi que deux modélisations de la relation dose-réponse (modèle logistique 4.1 et modèle puissance 4.2) ont été envisagés. Chaque simulation a généré 2 jeux de 10 000 essais indépendants correspondant respectivement à une première réponse fixée comme un échec ou à une première réponse fixée comme un succès. La probabilité ciblée était 0.1, la taille des cohortes de 1, le nombre de sujets inclus de 24. La première dose administrée était la plus faible dose et le saut de dose était autorisé (une dose peut donc être administrée sans que la dose juste en dessous n'ait été expérimentée). L'évaluation de l'impact de cette première réponse s'est faite sur la mesure du pourcentage de sélection correcte de la dose recherchée, le biais moyen des probabilités d'échec estimées, le nombre d'échecs observés et la répartition des doses administrées à la fin de l'essai.

Une approche utilisant une vraisemblance pondérée a été proposée afin de confirmer le rôle important jouer par les premiers individus inclus dans la MRS. Cette approche consiste à munir l'expression de la vraisemblance fournie par l'équation (4.3) de poids croissants avec le rang. En effet, comme nous l'avons écrit précédemment, l'information apportée par chaque observation est utilisée tout au long de l'essai. Il parait donc légitime de supposer que l'influence des observations est d'autant plus forte que l'observation est précoce. La fonction de vraisemblance pondérée après j observations s'écrit :

$$L_j(\theta) = \prod_{l=1}^{j} \psi(x(l),\theta)^{y_l w_{lj}} (1-\psi(x(l),\theta))^{(1-y_l)w_{lj}} \tag{4.8}$$

avec

$$w_{lj} = \frac{\log(\log(l+2))}{\sum_{i=1}^{j} \log(\log(i+2))}; l = 1,\ldots,j.$$

Le choix de la fonction log log permet de diminuer fortement le poids des premières observations et d'assurer des poids de même ordre de grandeur aux dernières. L'objectif est d'obtenir un algorithme relativement souple au début de l'essai sans trop perdre d'information lors de l'estimation finale.

Ce travail fait l'objet d'une soumission au journal *Clinical Trials*.

4.3.3 *Adaptive dose-finding designs for non cancer phase II trials : Influence of early unexpected outcomes*

Adaptive dose–finding designs for non cancer phase II trials: Influence of early unexpected outcomes

Matthieu RESCHE-RIGON, Sarah ZOHAR and Sylvie CHEVRET

Département de Biostatistique et Informatique Médicale; U717 Inserm; AP-HP; Université Paris 7; Hôpital Saint-Louis, 1 avenue Claude Vellefaux, 75475 Paris cedex 10, France.

Running head : Influence of early unexpected outcomes

Total number of words : 4946

Corresponding author :

Matthieu Resche-Rigon

Département de Biostatistique et Informatique Médicale

Hôpital Saint-Louis

1, avenue Claude Vellefaux

75010 Paris

Tel/Fax: (33) 142499773/ (33) 142499745

E-mail : matthieu.resche-rigon@univ-paris-diderot.fr

Abstract

In non cancer phase II trials, dose-ranging designs are usually based on fixed designs, where several doses, including a placebo, are randomly allocated to patients. However, in neonates or infants, there is an increased awareness of safety issues, avoiding randomization. We propose the use of adaptive designs such as the Continual Reassessment Method (CRM) that has been shown unbiased and efficient in cancer phase I trials. Based on a motivating example, we point out the individual influence of first outliers in this setting. Some method for the processing of outliers is proposed as a theoretical benchmark. Via simulations, we illustrate how this approach can provide us with further insight on the performance of CRM.

Keywords: Dose-finding, Individual influence, Continual Reassessment Method, Weights.

1 Background

Dose-finding clinical trials are characterized by small sample sizes, small sets of evaluated dose levels, and dose allocation schemes [1]. In life-threatening diseases such as cancer and AIDS, one cannot run a clinical trial with a new agent in patients without tolerance preoccupations [2], so that ethical constraints surround dose-finding designs. Therefore, doses escalate cautiously from a safe starting dose on the basis of current interim data. Previous dose levels and clinical outcomes are considered together to modify the dose allocated to the next new cohort of patients, resulting in adaptive designs. Several dose allocation schemes have been proposed, either algorithm– [3] or model–based [4, 5, 6, 7, 8, 9]. Among the latter, the Continual Reassessment Method (CRM) has been developed in a cancer phase I framework to estimate the maximal tolerated dose (MTD) of a new drug, which is taken to be the dose associated with a pre-specified "target" toxicity rate [4]. The CRM dose allocation rule is as follows. The first cohort of patients is administered either the initial MTD guess [4] or the first dose level [10, 11, 12, 13]. Once patient outcome, usually dose-limiting toxicity (DLT) or not, has been observed, the dose level assigned to the next patient cohort is that dose level associated with the estimated toxicity probability closest to the target. These probabilities are estimated from a parametric model, using either Bayesian (original CRM) [4, 10, 11, 12, 13] or likelihood (CRML) [14] inference. This is rerun until the pre-specified fixed sample size is reached.

By contrast, in the setting of phase II non cancer dose-ranging, trials take many forms and serve many objectives. When safety is not the main issue, due to previous reports in healthy volunteers, characterizing the dose-response relationship on the basis of phase II fixed designs is one of these objectives. Several doses are randomly allocated to

patients, throughout parallel or crossover designs, where a placebo (dose level zero) is often used as the control group. Instead of the MTD, dose response trials aim at estimating the minimum effective dose (MED), which is determined in comparison to the placebo, as the minimal dose producing an improvement of Δ [15]. Then, the problem of locating the MED is formulated as a multiple comparison problem. Nevertheless, in some phase II settings, ethical concerns are close to those observed in cancer phase I, so that minimizing the number of patients receiving non effective doses is mandatory. This is notably the case when dealing with infants or neonates, and though progress has been made in research on the effects of drug therapy on pediatric patients [16], neonates are still an understudied population where there is an increased awareness of safety issues [17].

To minimize the risks of clinical investigation in children, randomly allocating doses should be abandoned for improved dose adjustments [18]. Adaptive designs, with doses iteratively determined depending on data from prior dose levels, appear a promising issue for such trials. Actually, although the CRM(L) has been recommended for cancer phase I trials [19], it has been also used for non cancer phase II trials conducted in neonates [20, 21] and infants [22, 21], as well as in pregnant women [23]. In practice, doses are ordered by decreasing levels, and the dose-failure relationship is modeled directly [1]. This results in defining the MED as any targeted percentile of the dose-failure relationship. The main difference from former CRM(L) designs used in cancer phase I concerns the levels of target probability of failure as compared to that of toxicity. Actually, what constitutes acceptable probability of failure of drug is difficult to define, as compared to what defines acceptable probability of toxicity of most anticancer drugs. Indeed, since there is an implicit assumption of a positive correlation

between toxicity and efficacy of most cytotoxic drugs, some amount of toxicity is desired, usually between 20 and 33% [24]. By contrast, in non cancer phase II trials, reported target values of failure lie below 20%, down to 5% [22, 23, 20, 25, 21, 26]. Actually, the principal design objective of these trials is to estimate extreme percentiles of a dose-response distribution as precisely as feasible using the smallest number of experimental subjects possible. In such situations where prior opinion with regards to the failure probabilities lie between 0.01 up to 0.5, the impact of an early unexpected outcome on the allocated doses is likely to be high. Indeed, the more the probabilities are far away from 0.5, the more quantal-response models such as the logistic model are sensitive to small number of observations [27]. Such small sample properties have been also studied in the close setting of Robbins Monro process [28]. Actually, most of the information is carried out by the very few patients who experienced a response. Thus, failure observed in the first enrolled patient(s), who was administered the dose level associated with the lowest failure prior probability, which could be considered as an unexpected outcome (as it will be once in every 100 trials), is expected to be influential. This was illustrated in a real example from a phase II dose-finding trial in infants based on the CRM, where following a failure observed in the first infant, subsequent infants, all exhibiting a success, were administered the same dose level. This pointed out the issue of outliers in such dose-finding trials. The objective of this paper was thus to assess the individual influence of first outliers in this setting, using this trial as a real life case-study.

The paper is organized as follows. First, we present the example from real data that motivated this work, illustrating the individual influence of the first unexpected observation. Secondly, we studied whether this influence could be erased by using

weights relative to the rank in the likelihood. This is further assessed throughout a simulation study, and revisiting the example. Some concluding remarks are finally stressed out.

2 Motivating example

A phase II dose finding trial was conducted in a teaching-hospital from Paris area to assess the MED of Rocuronium, an aminosteroidal derived non-depolarizing neuromuscular blocking agent with a rapid to intermediate onset depending on dose, used to facilitate endotracheal intubation in infants (not published). Due to previous phase I reports, assigning doses within the range thought to be safe from pharmacokinetics and pharmacokinetic-dynamic modeling was allowed [29, 30], and safety was not the main concern of the trial that focused on estimating the MED. The MED was defined as the dose of Rocuronium required to allow easy or acceptable intubating conditions in 90% of infants. Such intubating conditions were assessed within 60 secondes following Rocuronium administration through the use of a previously published score [31]. The trial was designed according to the CRM, with a target 0.10 probability of failure, a fixed sample size of 25 patients, and using one patient cohorts. Six dose levels were studied, namely
$d_1 = 1.4,\ d_2 = 1.2,\ d_3 = 1.0,\ d_4 = 0.8,\ d_5 = 0.6,\ d_6 = 0.4\ mg.kg^{-1}$ with associated guesses of failure probability of 0.001, 0.05, 0.1, 0.3, 0.6, and 0.7, respectively. A logistic model was used as the dose-failure model, with intercept fixed at 3 [10, 11, 12, 32]. Inference was Bayesian, with unit exponential prior for model parameter. All these parameters were chosen according to a simulation study of model operating characteristics under several scenarios that were considered to encompass

what might happen, as commonly used in Bayes setting [8]. Estimated failure probabilities were based on the posterior mean of θ. Analyses were performed using the BPCT software [33].

The first infant, included in January, 2002, received the third dose level. Indeed, contrary to phase I settings where many authors have proposed to begin with d_1, in a phase II setting, it was more reasonable to begin with the initial MED guess d_3, since it was actually lower than d_1. The first infant experienced a failure. Thereafter, the first dose level was consecutively recommended from the 2^{th} to the 11^{th} infant, all exhibiting treatment successes. The trial ended in June, 2002 for technical reasons. Nevertheless, even if the trial accrual has been pursued, we computed that four more successful patients would have been included before recommending a decreased dose level for the 16^{th}. In other words, a total of 14 patients at the first dose, all successes, would have been required to modify the next allocated dose. Some could argue that such a finding results from the Bayesian inference of the CRM, related to the arbitrary choices for initial guesses of failure probabilities and for prior distribution of the model parameter. Thus, we reran analysis of the trial using CRML [14] without improving the results. Actually, a total number of 21 consecutive patients treated at d_1 and all exhibiting a success would have been required before modifying the recommended dose. In addition, at least seven additional successes at d_2 would have been required to experiment the target dose, d_3.

Based on this motivating example, we wondered whether the first failure could have been so influential on the dose allocation sequence, the reasons why, and how to erase such an influence.

3 Methods

Let $\Omega_j = \{x(i), y_i; i = 1, \ldots, j\}$, be the accumulated data after the inclusion of j patients, where $x(i)$ is the administered dose to the i^{th} patient, and y_i his(her) binary outcome (1=failure, 0=success).

Let $\psi(x_i, \theta) = P(y_i = 1|x_i)$ denote the one-parameter model for the dose failure relationship, where $\theta \in [0; \infty)$ is to be estimated, and $x_i = \psi^{-1}(p_i, \theta)$, where p_i is the initial guess of failure probability associated with the dose level d_i $(i = i, \ldots k)$ and θ_0 is the initial guess for θ. Two models were used, namely the mostly used logistic model, where the intercept a_0 is fixed at 3 [10, 12, 11, 32]

$$\psi(x_i, \theta) = \frac{\exp(a_0 + \theta x_i)}{1 + \exp(a_0 + \theta x_i)} \tag{1}$$

and the power model [14]

$$\psi(x_i, \theta) = x_i^{\theta}. \tag{2}$$

Note that $x_i = (logit(p_i) - a_0)/\theta_0$ and $x_i = \exp(\log(p_i)/\theta_0)$ for the logistic model and the power model, respectively.

The likelihood function is given by:

$$L_j(\theta) = \prod_{l=1}^{j} \psi(x(l), \theta)^{y_l} (1 - \psi(x(l), \theta))^{(1-y_l)} \tag{3}$$

Based on the Bayes estimate of θ with regards to the squared-error loss function, that is the posterior mean, the dose failure model is actualized, and that dose level associated with an estimated probability of failure closest to the target, p, is given to the next patient.

Since dose-finding sample sizes are small, small perturbations in dose allocation or responses are expected to modify the recommended dose level, resulting in individual influences [34]. Moreover, information drawn by any observation is used all along the

estimation process. Therefore, the influence of each observation is expected to be related to its rank, with the highest influence of first observations. Thus, to decrease the influence of the first observations, possibly outliers in response, weighted likelihood functions appear of interest [35, 36, 37, 38]. The weighted likelihood function was given by:

$$L_j(w_j, \theta) = \prod_{l=1}^{j} \psi(x(l), \theta)^{y_l w_{lj}} (1 - \psi(x(l), \theta))^{(1-y_l)w_{lj}} \quad (4)$$

By contrast to the likelihood CRM(L) given in equation (3), where all the weights w_{lj} ($j \leq n$) are equal to one, so that each patient contributes to the likelihood similarly whatever his(her) rank, we arbitrarily considered non unit weights, increasing with l and decreasing with j, as follows:

$$w_{lj} = \frac{\log(\log(l+2))}{\sum_{i=1}^{j} \log(\log(i+2))}; l = 1, \ldots, j$$

The choice of a log log shape was made to assure small weights only for first observations, and weights close to 1 otherwise. Of note, in case of cohort size above 1, the weights refer to the cohort rank rather than to the individual rank.
Estimation of the MED iteratively proceeded similarly to the CRM(L), that is using either Bayesian or likelihood inference, until the fixed sample size n is reached.

4 Simulation study

4.1 Simulated trials

We simulated $N = 10,000$ dose-finding trials, aiming at estimating, from 6 dose levels, the 10^{th} percentile of the dose-failure relationship. Six realistic scenarios were considered, with initial guesses of failure probabilities fixed at $0.01, 0.05, 0.1, 0.2, 0.3$ and 0.5 [14]. Scenario 1 reproduced the failure probabilities initially guessed in the

Rocuronium example. Scenario 2 supposed that the initial guesses described above were the underlying reality. In both cases, the MED was set at d_3. Otherwise, to assess the performance of CRM in case our prior expectations of the failure-dose curve have been too optimistic, the MED was set at d_2 in scenario 3 and 4, down to d_1 in scenario 5. Scenario 3 and 4 only differed in terms of failure probability at d_1, fixed at either 0.01 or 0.05, respectively. Finally, scenario 1* was similar to scenario 1, except that initial guesses of failure probabilities were set similar to reality, *i.e.*, at $0.001, 0.05, 0.1, 0.3, 0.6$ and 0.7. Both the logistic model previously used in the Rocuronium trial (equation (1) where $a_0 = 3$) and the power model (equation (2)) were fitted. A unit exponential prior for the model parameter, θ, was chosen for Bayes inference (thus, $\theta_0 = 1$). The posterior mean of θ was computed after each observation.

The patient cohort size was fixed at 1, and the fixed sample size at $n = 24$. The first dose level was administered to the first patient. Dose allocation scheme and inference used both the CRM and the CRML with skipping, with or without weights in likelihood. To recreate the real life case study that motivated this work, two settings were considered and compared. The first setting corresponded to trials where the first patient response was deterministically fixed at $y_1 = 1$ (*i.e.*, failure), while the second setting considered trials where $y_1 = 0$ (success). This did not aim at roughly comparing the overall performances of CRM and wCRM, since we were aware that both situations, that is $y_1 = 1$ and $y_1 = 0$ did not occur similarly in the real life, whatever the scenario, but rather to highlight the influence of y_1.

From the N independent replicates of each situation, operating characteristics were computed and compared, namely the percentage of correct dose selection (PCS) and the bias in estimated failure probability, both measured for the recommended dose level at

the end of the trial, as well as the average number of inclusions before ascending in dose levels, the average number of patients administered each dose level, and the overall failures in the trial.

Simulations were carried out in S language (R-cran 2.2 software). Code is available upon request to the first author.

4.2 Results

4.2.1 Unweighted likelihood for the CRM(L)

Table 1 represents the percentage of correct dose selection (PCS) at the end of the trial, using a logistic dose-response model, according to the response of the first patient, y_1. Similar findings are tabulated in Table 2 for the power model.

[Table 1 about here.]

[Table 2 about here.]

Considering the allocation of subjects based on CRM, the average proportions of failures for each of the six scenarios when $y_1 = 0$ were 0.117, 0.126, 0.127, 0.143, 0.148 and 0.178 for the power model. These failure proportions were slightly higher for the logistic model; they were above the target 0.10 in all the six scenarios. By contrast, when y_1 was fixed at 1, that is in case of a first observed failure, all failure proportions were below the target, ranging from 0.043 up to 0.097, except expectedly in scenario 5 where d_1 was the MED.

Figure 1 displays the number of allocated patients per dose level (left plots) when $y_1 = 1$, in each scenario. This confirms that, using CRM, only the first (in scenario 1^*) and the second dose levels were experimented once. As a result, the target dose level

was never administered (scenario 1*, 1, and 2) or only in less than 6 patients in 50% of cases (scenario 3) or in 25% of cases (scenario 4). This also explained why, due to the actually lower administered dose levels, the overall failures were lower as compared to the settings where $y_0 = 0$ as exposed above. Obviously, in scenario 5, where d_1 was the MED, all patients were administered this first dose level.

[Figure 1 about here.]

Nevertheless, in almost all situations, the PCS decreased when the first failure was fixed ($y_1 = 1$) as compared to a non-failure outcome ($y_1 = 0$). This was not related to the Bayesian estimation, and similar results were found using maximum likelihood estimator (data not shown). Note that the PCS reached by the use of the power model (Table 2) seemed always higher than those obtained with the use of the logistic model (Table 1). Finally, based on the CRM, a positive bias in estimating the failure probability of the MED was observed when $y_1 = 1$ (Table 1 and Table 2). This is also illustrated in the left part of Figure 2, that displays the failure probability at doses 1, 2, and 3, estimated through the use of the CRM after each inclusion based on one simulated set from Scenario 3, when $y_1 = 1$. True probabilities of failure were 0.01, 0.10 and 0.20, respectively. The overestimation of failure probabilities persisted all along the trial. This also illustrates why the administrated dose level was somewhat blocked on the first dose level.

[Figure 2 about here.]

4.2.2 Weighted likelihood for the CRM(L)

Using the weighted CRM allowed to widely modify these results. Though the PCS with constrained $y_1 = 1$ was decreased in most scenarios as compared to that reached with

$y_1 = 0$, the decrease was slighter than when using non weighted likelihood (Table 1 and Table 2). Similar results were observed with the maximum likelihood estimator (data not shown). Of note, when $y_1 = 0$, the PCS obtained with the weighted likelihood were close to or slightly worth than those obtained with the CRM.

The overestimation in failure probabilities previously observed with the CRM only persisted for the first allocations with the wCRM (Figure 2). Indeed, as compared to the CRM, when $y_1 = 1$, the wCRM allowed to experience an increased number of dose levels and dose allocation scheme rapidly "concentrated" around the target dose level, as expected. Moreover, in scenario 1^*, 1, and 3, the target dose level was the most frequently allocated dose level (Figure 1). As a result, the proportion of failures observed in case of $y_1 = 1$ were higher than those observed with the CRM (Table 1 and Table 2). Nevertheless, the average proportions of failures were close to those reached with CRM in case of $y_1 = 0$. Bias in failure probability estimates from the wCRM when $y_1 = 1$ were close to those obtained when $y_1 = 0$ (Table 1 and Table 2).

Otherwise, we similarly assessed the influence of second or later observation (data not shown). In these cases, the observed decrease in PCS with $y_i = 1$ compared to $y_i = 0$ persisted. Nevertheless, the difference in PCS obtained with fixed outcome of the i^{th} patient, $y_i = 1$ or $y_i = 0$, rapidly decreased according to the rank i.

5 Revisited exemple

5.1 Standard CRM

We first applied the power model $\psi(x_i, \theta) = x_i^{\theta}$ with unit exponential prior. Based on the observed patient outcome, this would have allowed the second dose level being

recommended after the sixth infant. The 7^{th} to 13^{th} infants would have been treated at the second dose level, and if all were still successes, the 14^{th} infant would have received the third dose level. This exhibits the improved performances of the power model over the logistic model, as observed from the simulation findings reported above.

In the setting of early phases trials, many authors have suggested that the CRM should begin with dose d_1. Thus, suppose that the first infant received the dose d_1. If he(she) experienced a failure, at least 58 infants with successes at d_1 would have been included before recommending d_2. By contrast, if the first infant experienced a success at dose d_1, then the second infant would have received the dose d_3 (power model) or d_4 (logistic model); If he(she) had a failure, the next recommended dose would have been the dose d_1, and this is as one would have expected in this situation.

Otherwise, prior choice, as well as the initial guesses of failure probabilities, could be influential in parameter estimation. We reran trial estimation using lognormal prior with mean 1 and decreasing variances. Expectedly, the lower the variance (*i.e.*, the stronger the prior), the lower the number of successes required to experience the second dose level.

Finally, the use of cohorts of 3 patients has been also proposed for CRM. We then assessed whether such a grouped CRM would have modified the results. Assuming that the second and third infants still exhibited successes at d_3, we found that the influence of the first unexpected outcome within the first cohort was erased, and that dose escalated early in the trial (Table 3). However, observations were quite modified, since we assumed that two further patients received d_3 in the first cohort.

[Table 3 about here.]

5.2 Weighted CRM

Downweighting the observations according to the rank through the use of weighted likelihood defined in equation (4), allowed confirming how the influence of the first observation from the Rocuronium trial was erased. Indeed, using the logistic model, the second dose d_2 would have been recommended for the third patient, while the third dose d_3 would have been the estimated MED since the sixth patient.

5.3 Bayesian decision procedures

We used the ADEPT software (SAS version 9) to perform the Bayesian decision procedures based on logistic regression models proposed by Whitehead and Williamson for dose-finding studies [39]. Based on the 6 dose levels available for testing in the Rocuronium trial, we attempted to retrospectively apply these methods using the observed doses and outcomes. The main target success probability was 0.9. It was believed that $TD90$ would be equal to $d_{-1} = 1.0$, and $TD50$ to $d_0 = 0.6$. Two forms of prior information were investigated, with pseudo-observations envisaged at dose d_{-1} and d_0, with the choice $n_{-1} = n_0 = 3$ (strong prior), or $n_{-1} = n_0 = 1$ (low prior). Based on patient gain function, which appears the closest allocation rule to that of CRM, the strong prior allowed to escalate to dose d_2 since the fourth patient, and to reach dose d_3 in the 14^{th}. The use of a low prior achieved an escalation to dose d_2 since patient 6, while escalation was still stopped after 30 simulated successes at that dose level. The results from the variance gain were somewhat different, with escalation at dose d_2 at the third (strong prior) or fifth (low prior) patient, while escalation reached dose d_3 at the seventh (strong prior) or tenth (low prior) patient. Actually, this decision-procedure approach appears to be less sensitive than the original CRM to the unexpected outcome

observed in the first patient.

6 Limitations

The aim of this paper was to illustrate the use of the CRM(L) for dose–finding experiments conducted in infants phase II trials. In this setting, ethical concerns are important, so that usual fixed designs may appear as a limit for inclusion. Actually, approximately 50% to 75% of drugs used in pediatric medicine have not been studied adequately to provide appropriate labeling information [16]. Therefore, the CRM, as an example of adaptive designs, could be a promising alternative approach.

We restricted the paper on phase II dose response studies, where estimating extreme percentiles of a dose-response distribution is the principal design objective, focusing attention on the cases that are most likely to be problematic, namely unexpected responses at the first administrated dose level. Indeed, owing to the very low failure probabilities guessed at first dose levels, it is expected that patients who receive such doses will not experience failure. For instance, using a 0.001 probability of response, it is expected that 999 individuals with non response could be observed before observing one response; this drops down to only 19 patients in case of a 0.05 probability. Thus, failures at first dose levels could be considered as occurrences of outliers, that is observations surprisingly extreme, that appear inconsistent with the rest of the data [40]. This was motivated by a real phase II trial in infants, where, unexpectedly, the first infant experienced failure. Contrary to phase I trials, where unexpected toxicity at the first dose level would require stopping the trial to better analyze such an occurrence and even to redefine the dose scale, initial failures are more acceptable in phase II trials, as illustrated above in the motivating example. Moreover, since all the next patients

experienced success, one could eliminate that the dose range was wrongly chosen. Therefore, this first observation could be considered as an outlier.

To get further insights in the CRM, initial outliers were investigated carefully. This was not done for practical implementation, but only as a theoretical benchmark. Actually, as shown in the simulation study, the CRM appears unable to rapidly overcome an early unexpected outcome, so that only the two first dose levels were experienced thereafter along the trial (Figure 1). This achieved a zero pourcentage of correct dose selection (PCS) in the cases where the target level was above these two first levels, *i.e.*, in scenarios 1, 3 and 4 (Table 1 and Table 2). Influence persisted whatever the inference process (Bayesian or maximum likelihood), the underlying dose–response model (logistic or power model) – though the power model performed better than the logistic model –, and the size of the cohort. Of note, in the wide literature devoted to CRM, these poor performances were neither reported. This could be explained by the usual settings of Monte Carlo studies assessing operating characteristics of the CRM, where responses are randomly generated. Since we focused on unexpected outcomes, they were randomly rarely observed (from 0.001 up to 0.05 of cases in the scenarios used). Secondly, most of the studies were performed in a phase I framework with a target around 0.3, that is where it is likely that influence would be erased.

Most statistical approaches developed to quantify individual influence of outliers are based on case-deletion methodology [34, 41]. The main idea of these approaches is to delete one observation, to re-fit the model on the remaining subjects, and then to choose an appropriate metric to measure changes in the estimator or any other statistics [42]. In the setting of dose-finding trials, such an individual influence should be estimated by comparing the MED estimate based on all patient data with that would have been

reached if one patient has not been enrolled. Since we cannot observe what might have happened, allocated doses being related to previous outcomes, these methods do not appear well suited to dose-finding trials. Thus, we evaluated the influence of first failure by comparing operating characteristics of CRM(L), constraining the patient response to 1 or 0. Since the influence of each observation in CRM is expected to be related to its rank, information being drawn by any observation is used all along the estimation process, we used weighted likelihood with weigths decreasing with the rank. We showed that downweighting the observations according to their rank allowed erasing such an influence. The weights were arbitrarily chosen, and many other weights could have been used. Once again, this was not the aim of this paper to definitely propose a weighted CRM(L) approach to cope with all outliers but rather to confirm that the influence of a first outlier is quite heavy.

We also considered some of the many alternatives to the CRM, to determine whether they run into similar problems. We shown that the use of cohorts of 3 patients, or of strongest prior, should have modified this first influence. The Bayes decision procedures proposed for dose-finding appeared less influenced by a first unexpected outcome [39]. Otherwise, one could suggest that early sensitivity could be overcame by using a two-stage procedure such as a first up-and-down approach followed by CRM. In practice, such a procedure would be unrealistic in a trial with less than 30 patients: indeed, an up-and-down procedure requires at least a cohort size of 7 or 13 patients to target a 0.1 or 0.05 probability of response, respectively [43].

In summary, we showed that CRM(L) appears of interest in phase II dose-finding studies. In this setting, target quantiles are often below 0.20, and the design appears quite sensitive to first unexpected outcomes, avoiding the experimentation of other dose

levels. This should be kept in mind when designing such trials. Some design characteristics should be carefully chosen, such as prior and initial guesses of failure probabilities, as well as the underlying dose-response model. Finally, the data should be finally carefully examined, in order to analyze why these outliers occurred. This could allow to detect potential contaminants (arising from some other distribution).

References

1. Chevret S. Basic concepts in dose-finding. In *Statistical Methods for dose-finding experiments*, Chevret SE (ed). John Wiley & Sons: Chichester, 2006; pp 5–18.

2. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *Journal of Clinical Oncology* 2001; **19**:265–272.

3. Storer BE. Design and analysis of phase I clinical trials. *Biometrics* 1989; **45**:925–937.

4. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 1990; **46**:33–48.

5. Gatsonis C, Greenhouse JB. Bayesian methods for phase I clinical trials. *Statistics in Medicine* 1992; **11**:1377–1389.

6. Whitehead J, Brunier H. Bayesian decision procedures for dose determining experiments. *Statistics in Medicine* 1995; **14**:885–893; discussion 895–899.

7. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* 1998; **17**:1103–1120.

8. Thall PF, Russell KE. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics* 1998; **54**:251–264.

9. Gasparini M, Eisele J. A curve-free method for phase I clinical trials. *Biometrics* 2000; **56**:609–615.

10. Faries D. Practical modifications of the continual reassessment method for phase Icancer clinical trials. *Journal of Biopharmaceutical Statistics* 1994; **4**:147–164.

11. Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine* 1995; **14**:1149–1161.

12. Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A comparison of two phase I trial designs. *Statistics in Medicine* 1994; **13**:1799–1806.

13. Moller S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Statistics in Medicine* 1995; **14**:911–922; discussion 923.

14. O'Quigley J, Shen LZ. Continual reassessment method: a likelihood approach. *Biometrics* 1996; **52**:673–684.

15. Ting N. *Dose Finding in Drug development*. Springer: New-York, 2006.

16. Roberts R, Rodriguez W, Murphy D, Crescenzi T. Pediatric drug labeling. Improving the safety and efficacy of pediatrics therapies. *JAMA* 2003; **290**:905–911.

17. Giacoia G, Mattison D. Newborns and drug studies: the NICHD/FDA newborn drug development initiative. *Clinical Therapeutics* 2005; **27**:796–813.

18. Pons G, Treluyer J, Dimet J, Merle Y. Potential benefit of bayesian forecasting for therapeutic drug monitoring in neonates. *Therapeutic Drug Monitoring* 2002; **24**:9–14.

19. Center for Drug Evaluation and Research. New drug development and review. Technical report. FDA. http://www.fda.gov/cder/handbook/, 2002.

20. Desfrere L, Zohar S, Morville P, Brunhes A, Chevret S, Pons G, Moriette G, Rey E, Treluyer J. Dose-finding study of ibuprofen in patent ductus arteriosus using the continual reassessment method. *Journal of Clinical Pharmacy and Therapeutics* 2005; **30**:121–132.

21. Treluyer J, Zohar S, E R, Hubert P, Iserin F, Jugie M, Lenclen R, Chevret S, Pons G. Minimum effective dose of midazolam for sedation of mechanically ventilated neonates. *Journal of Clinical Pharmacy and Therapeutics* 2005; **30**:479–485.

22. Fabre E, Chevret S, Piechaud JF, Rey E, Vauzelle-Kervoedan F, D'Athis P, Olive G, Pons G. An approach for dose finding of drugs in infants: sedation by midazolam studied using the continual reassessment method. *British Journal of Clinical Pharmacology* 1998; **46**:395–401.

23. De Spirlet M, Treluyer J, Chevret S, Rey E, Tournaire M, Cabrol D, Pons G. Tocolytic effects of intravenous nitroglycerin. *Fundamental and Clinical Pharmacology* 2004; **18**:207–213.

24. Storer B. *Statistics in Clinical Oncology*. In . Marcel Dekker: New York, 2001; pp 73–91.

25. Lefrere F, Zohar S, Bresson J, Chevret S, Mogenet A, Audat F, Durand-Zaleski I, Ghez D, Dal Cortivo L, Piesvaux P, Cavazzana-Calvo M, Varet B. A double-blind low dose-finding phase II study of granulocyte colony-stimulating factor combined with chemotherapy for stem cell mobilization in patients with non-hodgkin's lymphoma. *Haematologica* 2006; **91**:550–553.

26. Carvalho J, Balki M, Kingdom J, Windrim R. Oxytocin requirements at elective cesarean delivery: a dose–finding study. *Obstetrics and Gynecology* 2004; **104**:1005–1010.

27. Copas J. Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society, Series B* 1983; **45**:311–354.

28. Wetherill G, Glazebrook K. *Sequential Methods in Statistics*. Chapman and Hall: London, 1986.

29. Wierda J, Meretoja O, Taivainen T, Proost J. Pharmacokinetics and phamacokinetic-dynamic modelling of rocuronium in infants and children. *British Journal of Anaesthesia* 1997; **78**:690–695.

30. Scheiber G, Ribeiro F, Marichal A, Bredendiek M, Renzing K. Intubating conditions and onset of action after rocuronium in young children. *Pediatric Anesthesia* 1996; **83**:320–324.

31. Grand S, Noble S, Wood A, Murdoch J, Davidson A. Assessment of intubating conditions in adults after induction with propofol and varying doses of remifentanil. *British Journal of Anaesthesiology* 1998; **81**:540–543.

32. Potter D. Phase I studies of chemotherapeutic agents in cancer patients: A review of the designs. *Journal of Biopharmaceutical statistics* 2006; **16**:579–604.

33. Zohar S, Latouche A, Taconnet M, Chevret S. Software to compute and conduct sequential bayesian phase I or II dose-ranging clinical trials with stopping rules. *Comput Methods Programs Biomed* 2003; **72**:117–125.

34. Cook RD, Weisberg S. *Residuals and Influence in Regression*. Chapman and Hall: London, 1982.

35. Hu F, Rosenberger WF. Analysis of time trends in adaptive designs with application to a neurolophysiology experiment. *Statistics in medicine* 2000; **19**:2067–2042.

36. Hu F. The weighted likelihood. *The Canadian Journal of Statistics* 2002; **30**:347–371.

37. Park C, Basu A, Lindsay BG. The residual adjustement function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. *Computational Statistics and Data Analysis* 2002; **39**:21–33.

38. Markatou M, Basu A, Lindsay B. Weighted likelihood estimating equations: the discret case with applications to logistic regression. *Journal of Statistical Planning and Inference* 1997; **57**:215–232.

39. Whitehead J, Williamson D. Bayesian decision procedures based on logistic regression models for dose-finding studies. *Journal of Biopharmaceutical Statistics* 1998; **8**:445–467.

40. Barnett V, Lewis T. *Outliers in statistical data*. John Wiley & Sons: Chichester, 1994.

41. Cook D. Assessment of local influence. *Journal of the Royal Statistical Society, Series B* 1986; **48**:133–169.

42. Pan J, Fang K, von Rosen D. Local influence assessment in the growth curve model with unstructured covariance. *Journal of Statistical Planning and Inference* 1997; **62**:263–278.

43. Ivanova A, Montazer-Haghighi A, Mohanty S, Durham S. Improved up-and-down designs for phase I trials. *Statistics in Medicine* 2003; **22**:69–82.

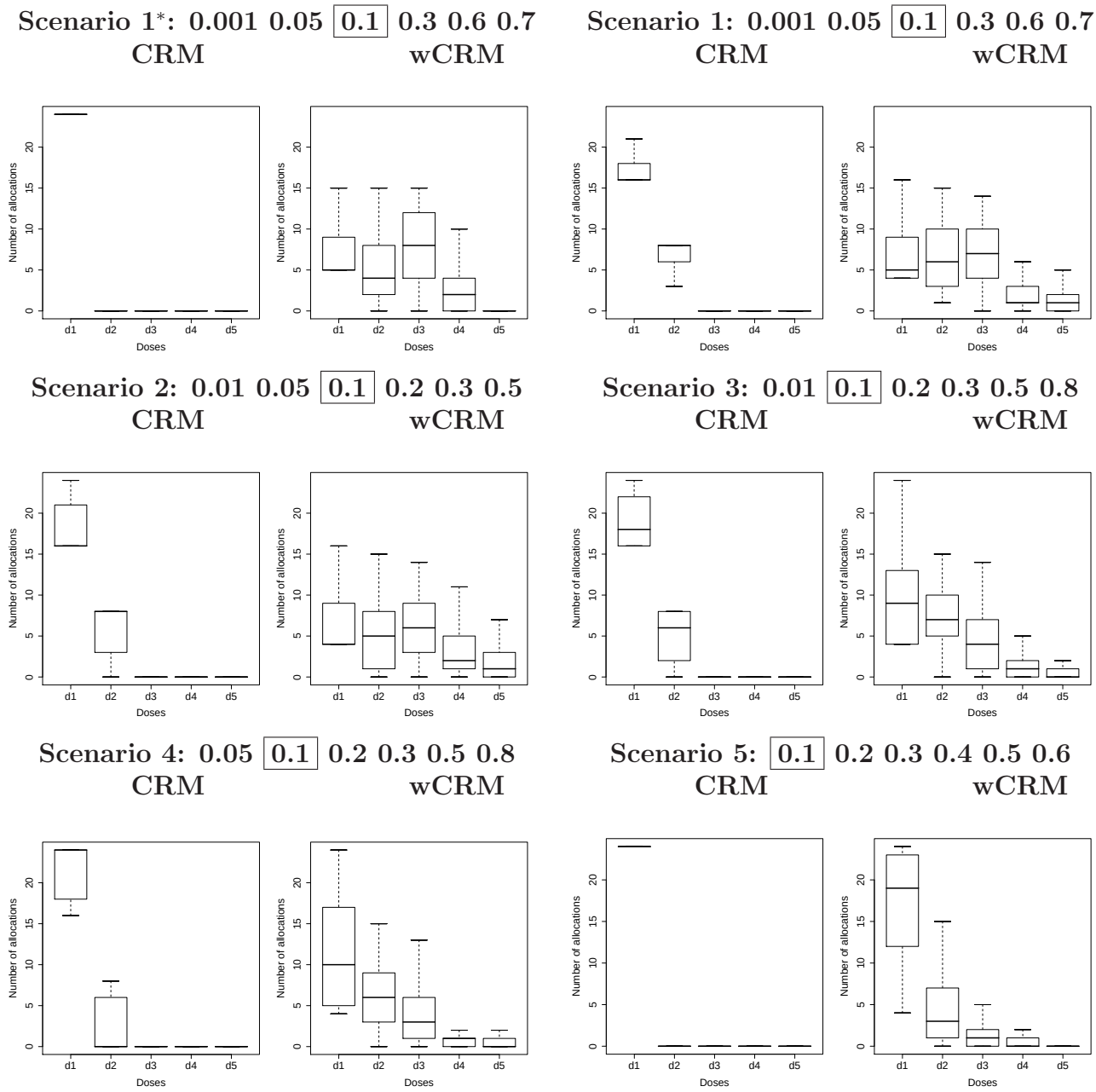


Figure 1: Number of allocations at dose levels d_1 to d_5 according to the likelihood (standard or weighted) among the $n = 24$ enrolled patients. For the 6 dose-failure scenarios, the first patient outcome was constrained to $y_1 = 1$, and cohorts of one patient per dose level were used. The underlying dose-failure model was a logistic model with intercept at 3. Scenarios used initial guesses of failure probabilities of 0.01, 0.05, 0.1, 0.2, 0.3, and 0.5, except scenario 1* that used 0.001, 0.05, 0.1, 0.3, 0.6 and 0.7. Target probability of failure was boxed (0.10).

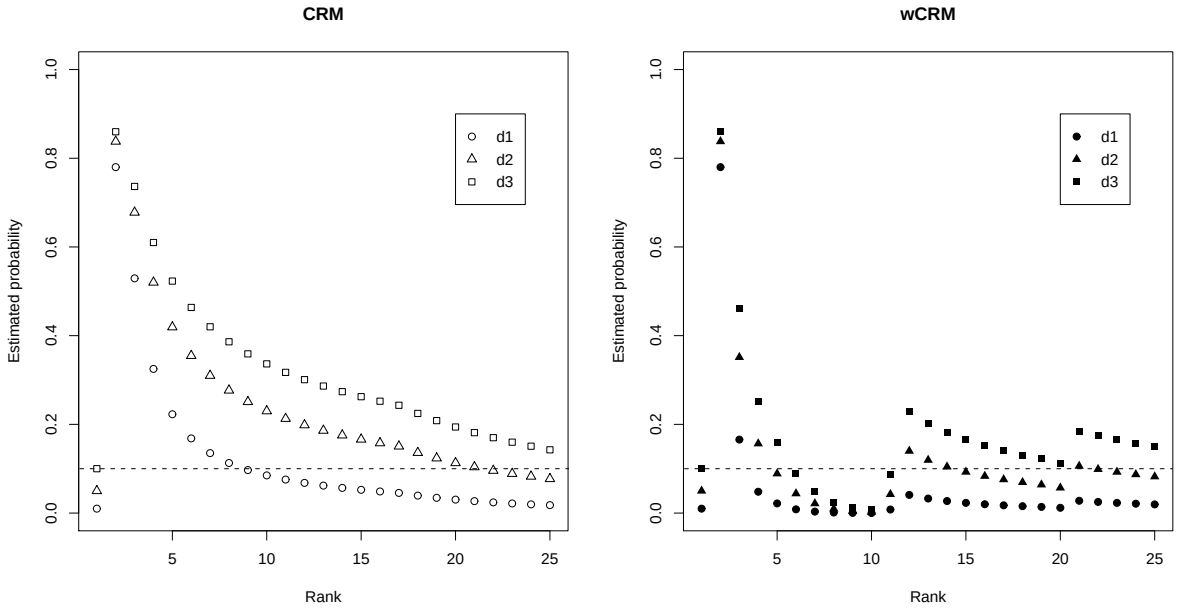


Figure 2: Plot of estimated failure probabilities derived from Bayes posterior mean of model parameter, at the three first dose levels d_1, d_2 and d_3, against the rank of included patients, according to the likelihood: unweighted (left plots) or weighted (right plots), using the most frequent simulated set from Scenario 3. The first patient outcome was constrained to $y_1 = 1$, and cohorts of one patient per dose level were used. The underlying dose-failure model was a logistic model with intercept at 3. The horizontal dashed line represents the targeted response probability, $p = 0.10$

Table 1: **Logistic dose-response model**: Percentage of correct selection (PCS), observed proportion of failures and median – absolute or relative – biases at the end of the trial, using either CRM or wCRM. In all scenarios, the first patient outcome was constrained ($y_1 = 1$ or $y_1 = 0$), and cohorts of one patient per dose level were used. Scenarios used initial guesses of failure probabilities of 0.01, 0.05, 0.1, 0.2, 0.3, and 0.5, except scenario 1* that used 0.001, 0.05, 0.1, 0.3, 0.6 and 0.7. Target probability of failure was boxed (0.10).

	CRM		**wCRM**	
	$y_1 = 1$	$y_1 = 0$	$y_1 = 1$	$y_1 = 0$
Scenario 1*: 0.001 0.05 [0.1] 0.3 0.6 0.7				
PCS	0.0000	0.5969	0.4861	0.5446
% Failure	0.0426	0.1325	0.1255	0.1415
Bias	0.0274	-0.0070	-0.0030	-0.0117
Relative Bias	27.4289	-0.0696	-0.0297	-0.1175
Scenario 1: 0.001 0.05 [0.1] 0.3 0.6 0.7				
PCS	0.0000	0.4682	0.4437	0.4696
% Failure	0.0563	0.1287	0.1444	0.1420
Bias	0.0271	0.0136	0.0186	0.0134
Relative Bias	0.5412	0.1355	0.1865	0.1341
Scenario 2: 0.01 0.05 [0.1] 0.2 0.3 0.5				
PCS	0.0000	0.4299	0.3785	0.4134
% Failure	0.0612	0.1440	0.1468	0.1479
Bias	0.0271	-0.0083	-0.0020	-0.0075
Relative Bias	0.5412	-0.0827	-0.0201	-0.0754
Scenario 3: 0.01 [0.1] 0.2 0.3 0.5 0.8				
PCS	0.5027	0.5275	0.5248	0.5142
% Failure	0.0698	0.1463	0.1555	0.1542
Bias	0.0528	0.0439	0.0466	0.0463
Relative Bias	2.0468	0.8786	1.0041	0.9416
Scenario 4: 0.05 [0.1] 0.2 0.3 0.5 0.8				
PCS	0.2694	0.4493	0.4206	0.4217
% Failure	0.0952	0.1522	0.1600	0.1592
Bias	0.0709	0.0482	0.0515	0.0498
Relative Bias	70.9157	1.1049	1.3199	1.137
Scenario 5: [0.1] 0.2 0.3 0.4 0.5 0.6				
PCS	0.9352	0.6987	0.7534	0.7111
% Failure	0.1415	0.1870	0.1863	0.1873
Bias	0.1137	0.0814	0.0892	0.0839
Relative Bias	113.7469	70.3889	83.7765	70.0110

Table 2: **Power dose-response model**: Percentage of correct selection (PCS), observed proportion of failures and median – absolute or relative – biases at the end of the trial, using either CRM or wCRM. In all scenarios, the first patient outcome was constrained ($y_1 = 1$ or $y_1 = 0$), and cohorts of one patient per dose level were used. Scenarios used initial guesses of failure probabilities of 0.01, 0.05, 0.1, 0.2, 0.3, and 0.5, except scenario 1* that used 0.001, 0.05, 0.1, 0.3, 0.6 and 0.7. Target probability of failure was boxed (0.10).

	CRM		**wCRM**	
	$y_1 = 1$	$y_1 = 0$	$y_1 = 1$	$y_1 = 0$
Scenario 1*: 0.001 0.05 [0.1] 0.3 0.6 0.7				
PCS	0.0000	0.6566	0.5691	0.6154
% Failure	0.0426	0.1165	0.1119	0.1158
Bias	0.0273	-0.0028	0.0023	-0.0124
Relative Bias	27.2679	-0.0277	0.0228	-0.1240
Scenario 1: 0.001 0.05 [0.1] 0.3 0.6 0.7				
PCS	0.0000	0.5363	0.484	0.5115
% Failure	0.0595	0.1256	0.0955	0.1296
Bias	0.0279	0.0077	0.0132	0.0058
Relative Bias	0.5583	0.0769	0.1322	0.0575
Scenario 2: 0.01 0.05 [0.1] 0.2 0.3 0.5				
PCS	0.0000	0.4607	0.4106	0.4482
% Failure	0.0637	0.1267	0.1329	0.1287
Bias	0.0279	-0.0110	-0.0016	-0.0098
Relative Bias	0.5583	-0.1095	-0.0162	-0.0983
Scenario 3: 0.01 [0.1] 0.2 0.3 0.5 0.8				
PCS	0.5933	0.5459	0.4861	0.5153
% Failure	0.0742	0.1430	0.1505	0.1444
Bias	0.0563	0.0423	0.0440	0.0447
Relative Bias	1.7723	0.845	0.8807	0.8936
Scenario 4: 0.05 [0.1] 0.2 0.3 0.5 0.8				
PCS	0.3369	0.4772	0.4160	0.4382
% Failure	0.0971	0.1482	0.1556	0.1491
Bias	0.0744	0.0458	0.0523	0.0490
Relative Bias	69.4047	0.9150	1.0733	0.9804
Scenario 5: [0.1] 0.2 0.3 0.4 0.5 0.6				
PCS	0.9094	0.6783	0.7143	0.6987
% Failure	0.1430	0.1776	0.1858	0.1870
Bias	0.1095	0.0821	0.0853	0.0835
Relative Bias	109.5305	70.3349	79.6169	73.2024

Table 3: Dose allocation rule for the Rocuronium trial using cohort of 3 patients per dose level (the 2^{nd} and the 3^{th} observations at d_3 are hypothetical

Cohort	Dose Given	No Failures	Posterior Estimate, θ	Estimated Failure Probabilities						Next Dose
				p_1^θ	p_2^θ	p_3^θ	p_4^θ	p_5^θ	p_6^θ	
CRM logistic										
1	3	1/3	0.695	0.020	0.244	0.352	0.581	0.768	0.818	1
2	1	0/3	0.772	0.009	0.169	0.266	0.507	0.730	0.792	2
CRM power										
1	3	1/3	0.608	0.015	0.162	0.247	0.481	0.733	0.805	2

4.4 MRS pondérée

Nous avons montré précédemment l'influence importante que pouvait prendre une première observation inattendue sur les résultats de la MRS. L'influence qu'exercent ces premières observations sur le résultat final semble essentiellement liée à l'impact qu'elles ont sur la séquence d'attribution des doses. Un résultat inattendu au départ semble "verrouiller" le processus sur la dose la plus faible (ou la plus élevée dans le contexte d'un essai de phase II).

Nous retrouvons donc les craintes exposées précédemment sur la MRS et déja observées avec la méthode de Robbins-Monro. Elles confirment l'opinion de Potter qui en 2006 [66] pointa le caractère moins robuste des méthodes paramétriques en comparaison des méthodes non paramétriques pour les essais de recherche de dose. En 2004, Roshan [82] proposa une modification de l'approche Robbins-Monro afin de remédier aux difficultés rencontrées pour l'estimation de quantiles extrêmes. Si les modifications ne sont pas transposables directement à la MRS, le concept développé par Roshan peut être retenu et appliqué à la MRS : "assurer un équilibre initial entre la montée et la descente potentielles des doses, puis se rapprocher d'un schéma classique". L'idée est donc d'agir essentiellement sur la séquence d'attribution des doses en rendant le schéma relativement souple au départ sans pour autant altérer l'estimation finale. Comme nous l'avons indiqué précédemment, le choix d'une dose se fait sur la base des observations précédentes, entraînant donc une dépendance des observations entre elles, même s'il existe une indépendance des réponses conditionnellement à la dose reçue. Nous proposons de tenir compte de cet aspect en utilisant le concept de vraisemblance pondérée "pertinente" (*Relevance weighted likelihood, ReWL*) développé par Hu, Rosenberger et Zidek [83, 84].

4.4.1 Vraisemblance pondérée pertinente *Relevance weighted likelihood*

Ce concept consiste à pondérer le poids des individus en fonction de leur pertinence [83, 84], c'est à dire de pouvoir apporter de l'information au modèle via la formulation de poids adaptés. Hu et Rosenberger [85] ont proposé d'utiliser cette méthode dans le cadre de schémas adaptatifs en présence d'une hétérogénéité temporelle. En effet, dans ce contexte (par exemple, du fait de la présence d'une phase d'apprentissage), l'hypothèse d'homogénéité de la distribution des réponses n'est plus valide. Ils proposent alors de diminuer le poids des premières observations en utilisant une fonction de vraisemblance pondérée, les poids étant des fonctions croissantes du rang des observations. La fonction de vraisemblance des observations $(x_i, i = 1, ..., n)$ s'écrit alors sous la forme suivante :

$$\prod_{i=1}^{n} f(x_i, \theta)^{w_{ni}} \tag{4.9}$$

où θ est le paramètre du modèle.

Les propriétés asymptotiques de la vraisemblance pondérée pertinente ont été démontrées par Hu en 1997 [83], sous réserve de vérifier deux conditions (i) $\forall i, w_i \geq 0$ et (ii) $\sum_{i=1}^{n} w_{ni} = 1$. Dans le cadre de l'estimateur du maximum de vraisemblance, la deuxième condition (ii) tombe, la localisation du maximum de la log vraisemblance ne dépendant pas d'un facteur multiplicatif. Dans leur article de 2000, Hu et Rosenberger [83] proposent ainsi deux fonctions de poids, polynomiaux : $w_{ni} = i^{\gamma}$, ou exponentiels : $w_{ni} = \gamma^{i}$, où γ est un paramètre positif à estimer en même temps que θ.

A la suite de nos précédents travaux, nous nous sommes proposés d'appliquer la vraisemblance pondérée pertinente à la MRS, l'objectif étant de diminuer l'impact des premières observations et d'améliorer la robustesse globale de la méthode.

4.4.2 Méthode de réévaluation séquentielle pondérée

Une fonction de vraisemblance pondérée a déjà été appliquée à la MRS. Ainsi, en 2000 Cheung et Chappell [86] ont proposé, dans le cadre d'essais de phase I où le délai d'obtention de la réponse est très long à observer, de pondérer les observations par une fonction dépendant de la durée de suivi des patients (*TITE-CRM*). En 2005, O'Quigley [87] proposa de même une MRS rétrospective où les observations étaient pondérées par la fréquence d'allocation des doses.

4.4.2.1 Vraisemblance pondérée pertinente pour la MRS

Nous nous sommes proposés d'appliquer la vraisemblance pondérée pertinente à la MRS utilisant l'estimateur du maximum de vraisemblance décrit par O'Quigley et Shen en 1996 [62].

Après j observations, la vraisemblance pondérée pertinente s'écrit alors :

$$L_j(\theta) = \prod_{l=1}^{j} \psi(x(l), \theta)^{y_l w_l} (1 - \psi(x(l), \theta))^{(1-y_l) w_l} \tag{4.10}$$

où $w_l = \log(\log(l+2))^{\gamma}$ et $\gamma \geq 0$.

Cette formulation de la vraisemblance est proche de celle de l'équation (4.8). Le choix de l'allure des poids est toujours dicté par la volonté de ne diminuer le poids que des premières observations.

L'estimateur du maximum de vraisemblance permet l'obtention de $\hat{\theta}_{|\Omega_j}$ et $\hat{\gamma}$. Il est intéressant de noter que, pour $\gamma = 0$, la vraisemblance pondérée pertinente est similaire à la vraisemblance non pondérée (equation 4.3).

4.4.2.2 Etude de simulation

L'intérêt de cette MRS modifiée a été étudié via la réalisation de simulations.

Plusieurs scénarios ont été envisagés, chacun étudié sur 20 000 essais indépendants. Pour chacun, la MRS utilisait un schéma *'3+3'* avant la première hétérogénéité des résultats, puis l'estimateur du maximum de vraisemblance pondérée pertinente (ReWL-CRM). Le saut de dose n'était pas autorisé et le modèle puissance (equation 4.2) décrivait la relation dose-réponse [62]. Enfin, diverses probabilités cibles (0.05, 0.1 et 0.3) ont été étudiées.

De façon à évaluer la robustesse de la méthode proposée, et étudier l'influence du rang des observations, les simulations ont été reconduites en introduisant des contaminants, patients pour lesquelles la relation dose-toxicité était décalée vers le haut. Ces contaminants ont été introduits en proportion constante (10%) dans chaque quart ordonné de l'échantillon.

Les résultats en termes d'estimation et de sélection correcte de dose sont améliorés par la ReWL-CRM d'autant plus que la cible s'éloigne de 0.30. Enfin, la ReWL-CRM est plus robuste aux contaminants précoces, illustrant indirectement que la méthode standard possède une influence importante aux premières observations.

Ce travail fait l'objet d'une soumission à *Statistics in Medicine*.

4.4.3 *Weighted Continual Reassessment Method*

STATISTICS IN MEDICINE

Statist. Med. 2006; :1–2 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Maximum Relevance Weighted Likelihood Estimator: Application to the Continual Reassessment Method

Matthieu RESCHE-RIGON*[1,2,3], Sarah ZOHAR[1,3] and Sylvie CHEVRET[1,2,3]

1 Département de Biostatistique et Informatique Médicale, Hôpital Saint-Louis, AP-HP, Paris, France; 2 Université Paris 7 - Denis Diderot, Paris, France; 3 Inserm, U717, Paris, France.

SUMMARY

Usual phase I dose-finding clinical trials, notably in cancer, are characterized by a small number of included patients (less than 40), a relatively high number of dose levels (4 to 6) and sequential dose allocation rules. In this setting, the Continual Reassessment Method (CRM) has been advised to be used routinely to provide a consistent and unbiased estimate of the maximal tolerated dose (MTD) of a new drug, possibly based on likelihood (CRML). In this adaptive design setting, we derived a Relevance Weighted Likelihood to propose a robust estimation of the MTD. The main idea is to weight the individual contributions to the likelihood using a decreasing function of the rank. We compare this method to the CRML throughout simulations. Copyright © 2006 John Wiley & Sons, Ltd.

*Correspondence to: Département de Biostatistique et Informatique Médicale, AP-HP, Hôpital Saint-Louis, 1 avenue Claude Vellefaux, 75475 Paris cedex 10, France. E-mail: matthieu.resche-rigon@paris7.jussieu.fr

1. INTRODUCTION

The primary objective of phase I dose-finding cancer trials is to estimate the maximum tolerated dose (MTD) of a new drug among a few number of dose levels. Traditional cancer phase I designs are rule-based designs that treat the MTD as a sample statistic. More recently, model-guided designs have been developed, where the MTD is considered as the dose that corresponds to a prespecified probability of toxicity in the patient population, that is, some percentile of interest. In most of these designs, an inference process on a parametric framework is used to guide dose escalation [1, 2, 3, 4, 5]. First of all, the Continual Reassessment Method (CRM) was proposed by O'Quigley *et al*. to estimate the MTD [1]. While originally based on a Bayes inference, a likelihood version of the original CRM (CRML) was thereafter developed [6]. This further requires a set of heterogeneous responses, so that the first stage is a rule-based design using three patient cohorts that ends upon observation of the first toxicity.

Whatever rule- or model-based, CRM(L) is an adaptive design where future design points are selected on the basis of previous responses at previous design points. Indeed, the dose administered to any patient is selected on the basis of previous design points, namely patient doses and responses. Because dose assignements depend on previous data collection, observations are dependent. Therefore, information drawn by any observation is used all along the estimation process. Thus, the influence of each observation is expected to be related to its rank, with highest influence of first observations. This influence could be reinforced by the small sample size of dose-finding trials. Moreover, as a binary regression, there is no response symmetry. The more estimated probabilities are far from 0.5, the more models are sensitive to a small number of observations [7]. Indeed, when estimated probabilities are lower than 0.5, most information is carried out by the very few patients who experiment a response. This is

well known in the setting of Robbins-Monro procedure that this does not perform well in the estimation of extreme quantiles [8]. In a phase I dose-finding cancer trial setting, the target probability usually ranges below 50%, commonly between 20% and 30% [1]. In phase II dose-finding trials that focus on probabilities of failure, targets are even more likely to lie about 10% or even 5% than about 30%. In the setting of such possibly extreme quantiles, the robustness of CRM(L) as a model-guided design has been pointed out as compared to rule-based designs, considered as "intrinsically robust" by Potter [9].

Actually, the hidden assumption of the CRML statistical modelling, *i.e.*, that the probability distribution of dose-response is homogeneous, can be violated [10]. In the setting of adaptive designs with time heterogenity, relevance weighted likelihood (ReWL) methods have been proposed by Hu and Rosenberger [10, 11]. They consist in weighting individual contributions to the likelihood according to their relevance, in order to decrease the influence of first observations on global conclusions [12, 13]. In our context of individual rank-related influence, we propose to develop a robust method for CRML, by weighting the individual contributions to the likelihood according to their rank, using ReWL. This could be easily applied to the Bayesian CRM.

The paper is organized as follows. First, we present the weighted estimator of ReWL. Then, Section 3 provides a simulation study to assess its relative performances as compared to standard likelihood CRML. Results are presented in Section 4. Finally, some discussion with practical considerations is provided in Section 5.

2. ROBUST LIKELIHOOD CONTINUAL REASSESSMENT METHOD

2.1. Continual Reassessment method

The CRM is based on a sequential estimation of the MTD from a finite set of dose levels and a fixed sample size. Let d_i $(i = i, \dots k)$ denote the dose levels of the drug to be tested and p the target probability of response. The relationship between dose and response is modeled through a one-parameter model $\psi(x_i, \theta)$, where θ is the parameter to be estimated and x_i a function of doses given by $\psi^{-1}(p_i, \theta_0)$, p_i being the initial guess of toxicity probability associated with the dose level d_i and θ_0 the initial guessed value of the parameter θ. In this work, we used the power model $\psi(x_i, \theta) = x_i^{\exp(\theta)}$ with $\theta_0 = 0$ as described by O'Quigley and Shen [6]. Since $\theta_0 = 0$ is chosen, the ψ function then reduces to $\psi(x_i, \theta) = p_i^{\exp(\theta)}$.

Let $\{(x(r), y_r);\ r = 1, \dots, j\}$ be the accumulated data after the inclusion of the j^{th} patient, with $j \leq n$, the fixed sample size, $x(r)$ the administered dose to the r^{th} patient, and y_r his (her) binary outcome.

The likelihood function $L_j(\theta)$ after j patients is defined by:

$$L_j(\theta) = \prod_{r=1}^{j} \psi(x(r), \theta)^{y_r} (1 - \psi(x(r), \theta))^{(1-y_r)} \tag{1}$$

Attribution of doses is iteratively performed after each observation by the selection of the dose level $x(j+1)$ which minimizes $(\psi(x_i, \hat{\theta_j}) - p)^2; i = 1, \dots k$ where $\hat{\theta_j}$ is the updated model parameter through maximum likelihood estimation [6]. This will be referred as the CRML below.

To reduce the impact of first observations, we proposed to adapt weighted likelihood estimators such that proposed by fHu and Rosenberger [10, 11] or the CRML. Each individual component of the likelihood (1) is thus weighted differently, so that the likelihood after j

patients becomes :

$$L_j^w(\theta) = \prod_{r=1}^{j} \psi(x(r),\theta)^{y_r w_r} (1-\psi(x(r),\theta))^{(1-y_r)w_r} \quad (2)$$

where w_r is the weight of the r^{th} out of j patients. To slowly increase weights over ranks, the weight w_r of the r^{th} included patient was defined as follows:

$$w_r = \log(\log(r+2))^{\gamma} \quad (3)$$

with $\gamma \in [0, 4.5]$. After j included patients, $\hat{\theta}_j$ and $\hat{\gamma}_j$ are estimated by maximization of the weighted likelihood. The administrated dose level to the next patient is that dose level associated with the estimated probability of response closest to the target. This allocation procedure will be further denoted ReWL CRM.

3. SIMULATION STUDY

We simulated $N = 20,000$ phase I cancer dose-finding trials aiming at estimating the 10^{th} percentile of the dose-toxicity relationship. Six dose levels were considered, with initial guesses of toxic probabilities (so-called working model) fixed at 0.01, 0.05, 0.1, 0.2, 0.3 and 0.5, respectively.

Six scenarios of actual toxic probabilities, $Sc_s(x) = P(Y = 1|x)$ $(s = 1, \ldots 6)$ were examined (Figure 1). In scenario 1, the actual probabilities are equal to the working model. In scenario 2, the first dose level is noticably nontoxic (10 fold lower than in scenario 1). Scenario 3 is similar to scenario 2, although the rate of toxicity above the MTD is greater than in scenario 2. Scenario 4 and 5 are close to scenario 3, though with increased toxicity since the first dose level. Finally, in scenario 6, toxicity is noticably excessive for all doses. Note that, in scenario 3, the differential in toxicity between doses around the MTD is higher than in scenarios 1-2

and 4-5, so that the MTD will be simplier to identify.

[Figure 1 about here.]

The trial sample size was fixed at $n = 24$. The first dose level was administered to the first patient. Dose allocation scheme and inference used the standard CRML and ReWL CRML, unless no heterogeneity in responses where the standard '3+3' scheme was first used, similarly to the CRML design of O'Quigley and Shen [6]. No skipping was allowed.

To better assess the robustness of the method, that is to highlight the rank influence responsible for somewhat dose-response heterogenity, we simulated a contaminant population according to the rank. Briefly, highly toxic contaminants were generated from $Sc_s(x)^{\exp(\beta)}$ where $\beta = -2$, except for scenario 6 where $\beta = 2$, and concentrated within each quarter of the sample. Overall proportion of contaminants was fixed at 0.10, so that the expected number of contaminants was 2.4, all observed within each quarter of the sample, that is, within each subset of 6 patients.

To confirm the increased individual influence in case of low target levels, we reran simulated trials using a 0.05 target, from the first three scenarios, where the MTD was the second dose level. Finally, to asses the performances of the method when dealing with higher target levels, we reran analyses using 0.30 as the target.

In each situation, operating characteristics were computed and compared from the 20,000 simulated trials, namely the percentage of dose correct selection (PCS) and the estimated response probabilities with mean bias and mean squared error (MSE), and the overall toxicities in the trial.

Simulations were carried out in S language (R-cran 2.2 software). Code is available upon request to the first author.

4. RESULTS

4.1. Comparison of CRML and ReWL-CRM according to the target

Table I reports simulation results when dealing with a target probability of 0.30, 0.10 and 0.05. When dealing with a targeted 30^{th} percentile of the dose-toxicity relationship, performances of CRML and ReWL CRM were close either in terms of PCS, bias and MSE or in terms of observed toxicities. However, when the target was 0.10, PCS were improved while biases and MSE decreased by using ReWL CRM as compared to the use of CRML. This was further observed when dealing with a 0.05 target, where gain in PCS achieved by the use of ReWL CRM reached about 10% in the three scenarios. These findings illustrate how, as stated above, individual influence in CRM is related to the rank, with heavier influence in case of low targeted percentiles.

[Table 1 about here.]

4.2. Increased heterogenity in dose-response

To better exemplified how the ReWL CRM outperforms, that is by erasing the first individual influence as compared to CRML, we further simulated a contaminant population within each quarter of the sample. Results are displayed in Figure 5 when dealing with a 0.10 target of toxic probability. Actually, earlier were the contaminants, higher was the difference in PCS and bias from the two methods, with improved performances of the ReWL CRM. This confirms that the ReWL CRM erases the influence of first observations, which is obvious in the CRM.

[Figure 2 about here.]

5. DISCUSSION

It has been established that the CRM is consistent under model misspecification but not generally. This paper pointed out the rank influence in the CRM(L) when estimating the MTD, which was assessed throughout a simulation study. We also wondered whether the robustness of the CRM could be improved by downweighting the influence of first observations. Indeed, since observations are made sequentially by the dose-finding design, the probability distribution of the responses has been reported potentially time heterogeneous [10]. In such a setting, the potential for time trends could bias results from standard likelihood analyses and it is desired that weighted likelihood methodology be used to take this time trend into account. Actually, results of our simulation study showed the superiority of the ReWL CRM over CRML with respect to both the correct estimation of the MTD and accuracy, especially in the cases where MTD was defined as a low percentile of the dose-toxicity relationship. Indeed, when the percentile of interest was low, the ReWL CRM was less sensitive than CRML in terms of how close the converged recommendation is to the target. Moreover, both the bias and the mean square error were reduced, in agreement with previous reports from another settings [10]. Finally, the ReWL estimator depends on the relevance weights, that express the statistician's perceived relationship among the studied population, and are usually chosen on intuitive grounds [13]. Accordingly we decided to choose a decreasing function of the rank that mostly affects the first ranks.

Several robust methods have been proposed for linear regression, then modified for the logistic model. These methods, such as M-estimator [14, 15] and E-estimator [16], consist in downweigthing observations with large residuals at the time of analysis. When analysis is performed sequentially, this is questionable. Notably, detecting "large" residuals in comparison

to the others is an open issue. Thus, we retained the widest concept of weighted likelihood that better handles the sequential nature of the CRML. The theory of weighted likelihood has been used in a diverse group of applications. Actually, it has been already used in the setting of CRM [17, 18]. O'Quigley developed the so-called "retrospective CRM" to re-analyse dose-findings trials through the CRM by weighting observations with the frequency of previous dose allocation [17]. Cheung proposed the time-to-event CRM (TITE-CRM) that allows patients to be entered in a staggered fashion, with weights depending on the time-to-analysis [18]. We used a decreasing function of the rank over a bounded interval, to insure a potentially heavy decreased influence of first observations but close weights for the last ones. Of note, when $\gamma = 0$, weights were all equal to one, so that the weighted likelihood (2) is equivalent to the standard likelihood (1).

Finally, we recommend the use of ReWL CRM. Although a simulation study cannot represent the universally valid truth in a strongly mathematical sense, it allowed learning about the properties of the design in various situations. Most scenarios actually included the true MTD, but we also considered an extreme scenario, that investigated a dose range completely located over the true MTD. In all situations, ReWL CRM performs better than the standard CRML in terms of PCS, and as efficient as the standard CRML from ethical points of view. This should not avoid the classical rules of prudence when conducting dose-finding trials: Treating patient one by one (even with cohort size greater than one), including patient once previous patient's response has been observed, and sequentially computing stopping rules based on toxicity to avoid misconclusions [19, 20].

REFERENCES

1. J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*, 46:33–48, 1990.
2. C. Gatsonis and J. B. Greenhouse. Bayesian methods for phase I clinical trials. *Statistics in medicine*, 11:1377–1389, 1992.
3. J. Whitehead and H. Brunier. Bayesian decision procedures for dose determining experiments. *Statistics in medicine*, 14:885–893; discussion 895–899, 1995.
4. J. Babb, A. Rogatko, and S. Zacks. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in medicine*, 17:1103–1120, 1998.
5. P. F. Thall and K. E. Russell. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*, 54:251–264, 1998.
6. J. O'Quigley and L. Z. Shen. Continual reassessment method: a likelihood approach. *Biometrics*, 52:673–684, 1996.
7. J.B. Copas. Binary regression models for contaminated data. *Journal of the royal statistical society. Series B*, 50:225–265, 1988.
8. V. Roshan Joseph. Efficient Robbins-Monro procedure for binary data. *Biometrika*, 91:461–470, 2004.
9. D.M. Potter. Phase I studies of chemotherapeutic agents in cancer patients: A review of the designs. *Journal of Biopharmaceutical statistics*, 16:579–604, 2006.
10. F. Hu and W. F. Rosenberger. Analysis of time trends in adaptive designs with application to a neurolophysiology experiment. *Statistics in medicine*, 19:2067–2042, 2000.
11. F. Hu and W. F. Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials*, chapter 7, pages 107–119. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, New Jersey, 2006.
12. F. Hu and J.V. Zidek. The weighted likelihood. *The Canadian journal of statistics*, 30:347–371, 2002.
13. F. Hu. The asymptotic properties of the maximum relevance weighted likelihood estimators. *The Canadian journal of statistics*, 25:45–49, 1997.
14. F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics*. Wiley, New York, 1986.
15. M.P. Victoria-Feser. Robust inference with binary data. *Psychometrika*, 67:21–32, 2002.
16. A.F. Ruckstuhl and A.H. Welsh. Robust fitting of the binomial model. *The annals of statistics*, 29:1117–1136, 2001.
17. J. O'Quigley. Retrospective analysis of sequential dose-finding designs. *Biometrics*, 61:749–756, 2005.

18. Y. K. Cheung and R. Chappell. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*, 56:1177–82, 2000.

19. J. O'Quigley and E. Reiner. A stopping rule for the continual reassessment method. *Biometrika*, 85:741–748, 1998.

20. S. Zohar and S. Chevret. The continual reassessment method: Comparison of bayesian stopping rules for dose-ranging studies. *Statistics in medicine*, 20:2827–2843, 2001.

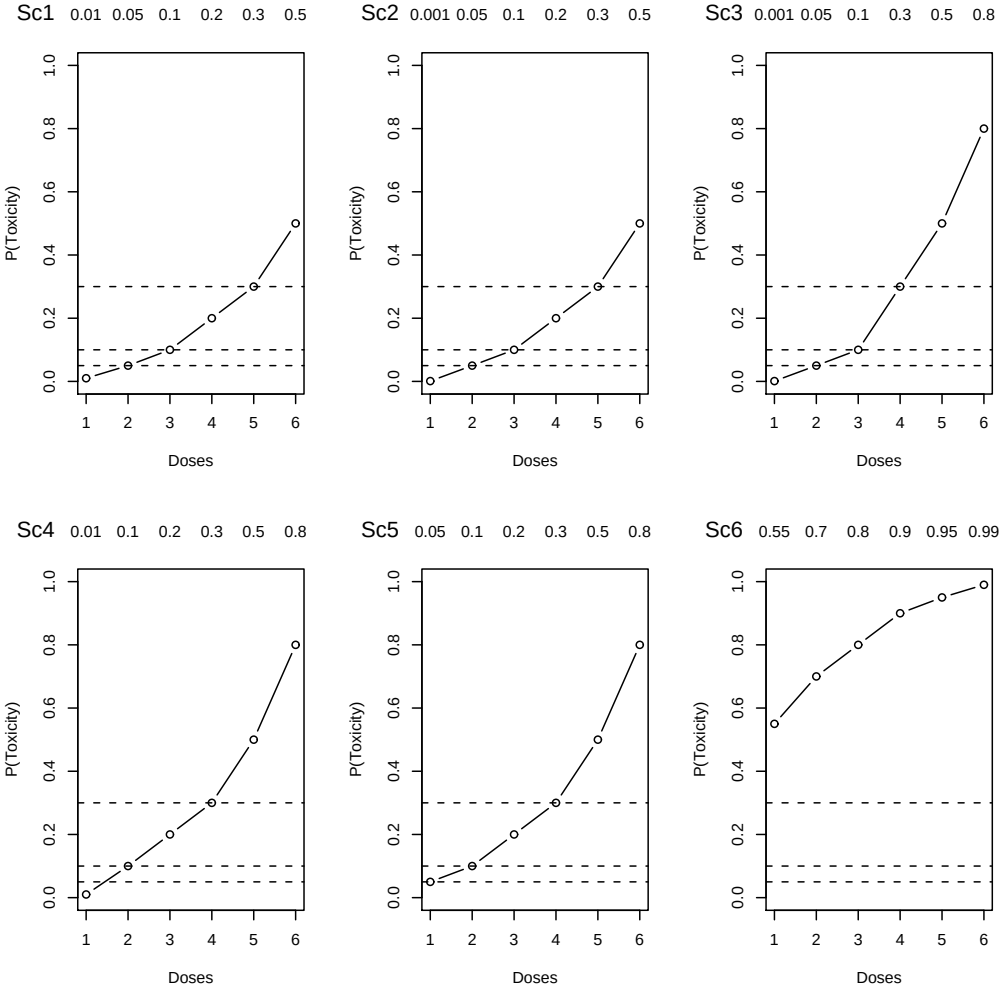


Figure 1. Dose-toxicity curves: Scenarios 1 to 6.

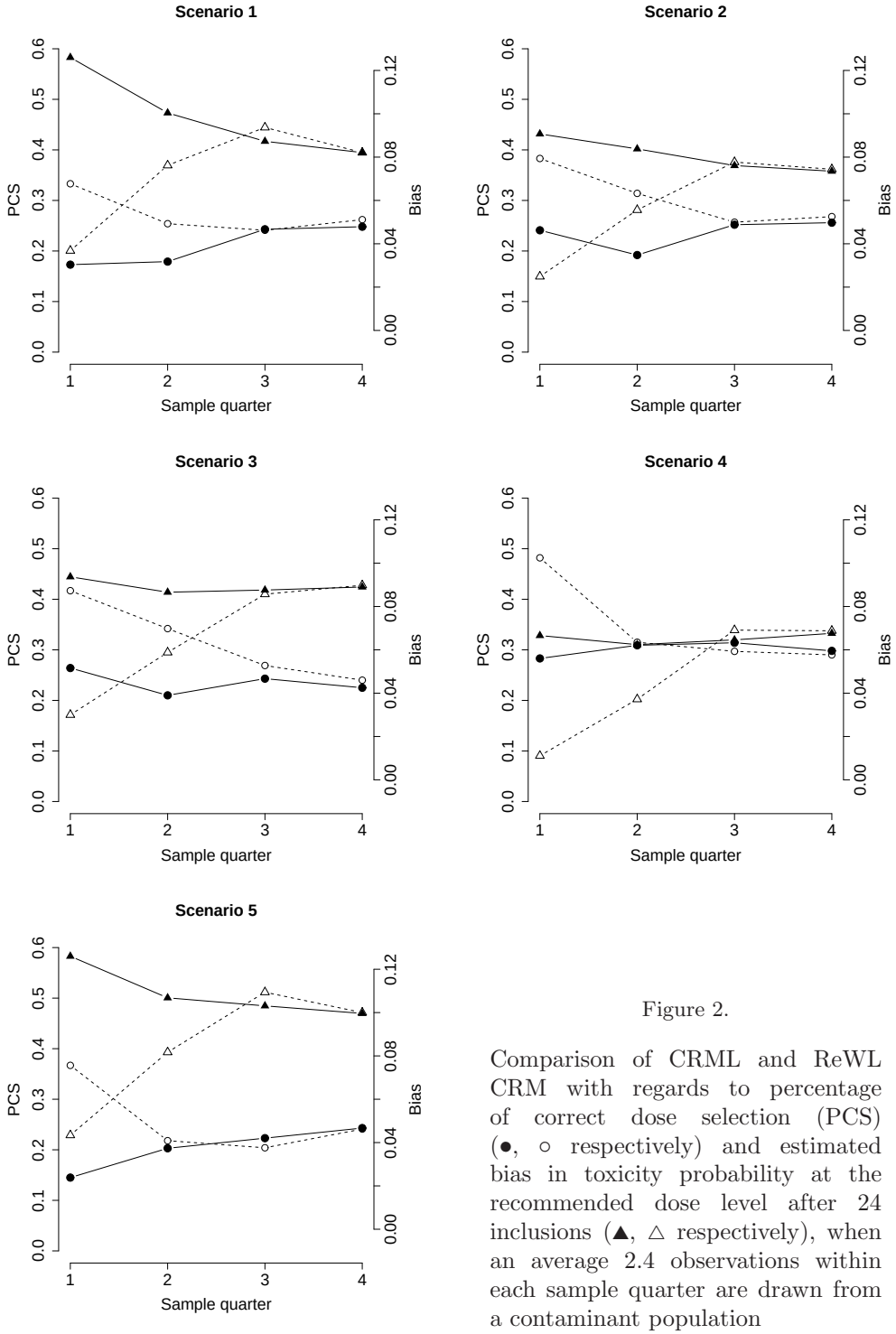


Figure 2.

Comparison of CRML and ReWL CRM with regards to percentage of correct dose selection (PCS) (●, ○ respectively) and estimated bias in toxicity probability at the recommended dose level after 24 inclusions (▲, △ respectively), when an average 2.4 observations within each sample quarter are drawn from a contaminant population

Table I. Comparison of the Standard CRML or the ReWL CRM when estimating the MTD defined as the 5th, 10th or 30th percentile of the dose-toxicity relationship, based on 24-patient cohorts and 20,000 replicated trials from the six different scenarios (Sc)

Target	Sc.	Method	Dose recommendation (%)			Bias	MSE	% toxicities
			Too low	MTD	Too high			
0.30	1	CRML	27.9	52.2	19.9	-0.010	0.098	19.8
		ReWL	25.5	53.8	20.7	-0.016	0.098	20.0
	2	CRML	26.6	53.0	20.4	-0.013	0.098	20.0
		ReWL	25.1	54.0	20.9	-0.018	0.098	19.9
	3	CRML	10.6	60.0	29.4	-0.019	0.068	24.9
		ReWL	8.9	59.7	31.5	-0.024	0.069	24.1
	4	CRML	24.8	53.1	22.0	0.009	0.080	24.5
		ReWL	23.1	53.4	23.5	0.005	0.080	24.1
	5	CRML	29.1	51.1	19.8	0.018	0.085	23.9
		ReWL	25.4	52.2	22.5	0.010	0.084	24.4
	6	CRML		99.9	0.1	0.001	0.102	55.4
		ReWL		99.8	0.2	0.003	0.113	55.8
0.10	1	CRML	30.1	41.3	28.6	0.007	0.065	9.5
		ReWL	27.5	43.8	28.7	-0.000	0.061	9.4
	2	CRML	27.7	42.9	29.4	0.002	0.060	9.6
		ReWL	26.8	45.0	28.2	-0.002	0.057	9.4
	3	CRML	36.8	49.2	14.1	0.059	0.179	13.1
		ReWL	30.0	55.6	14.4	0.010	0.052	9.7
	4	CRML	19.5	46.2	34.3	-0.002	0.049	10.7
		ReWL	18.6	46.4	35.0	-0.006	0.050	10.6
	5	CRML	34.8	38.5	26.7	0.025	0.076	11.2
		ReWL	29.2	39.4	31.4	0.013	0.071	11.3
	6	CRML		100.0		0.001	0.102	55.2
		ReWL		100.0		0.003	0.113	55.2
0.05	1	CRML	36.4	36.4	27.2	0.011	0.046	7.0
		ReWL	26.3	45.7	28.0	0.007	0.043	6.8
	2	CRML	33.1	38.9	27.9	0.004	0.035	6.8
		ReWL	21.7	49.5	28.7	0.002	0.034	6.6
	3	CRML	36.6	40.4	23.0	0.008	0.035	6.7
		ReWL	24.9	50.9	24.2	0.006	0.034	6.5

4.5 Conclusions

Comme nous le soupçonnions et comme certains auteurs l'avaient montré pour la méthode de Robbins-Monro [80, 81, 82], les performances de la MRS pour la recherche de faibles quantiles sont décevantes. Cela peut s'avérer problématique lors de l'utilisation de la MRS pour des essais de phase II lorsque les probabilités cibles sont de l'ordre de 10%, comme l'ont illustré les résultats observés sur un exemple réel. Nous avons montré par simulation que cette constatation est liée en partie à l'influence importante de la première observation et, plus généralement, des premières observations lors du déroulement d'un essai conduit avec la MRS. L'observation précoce d'un résultat peu attendu conduit l'essai à rester focalisé sur certaines doses, sans parvenir à expérimenter suffisamment d'autres doses parmi les sujets de l'essai, en nombre prédéterminé.

Cette influence semble peu liée au choix de l'inférence puisque les résultats obtenus par simulation sont similaires que l'on utilise la MRS bayesienne ou basée sur le maximum de vraisemblance. De même, la famille de la relation dose-réponse choisie ne modifie que peu les résultats.

Finalement, nous avons montré que la nature séquentielle de la MRS explique en grande partie cette influence. En effet, l'information apportée par une observation est réutilisée à chaque attribution ultérieure de doses, c'est à dire, à chaque nouvelle estimation des paramètres de la relation dose-réponse. Nous avons ainsi montré que diminuer de manière fixe l'impact des premières observations à l'aide d'une vraisemblance pondérée conduit à fortement diminuer l'impact d'une observation inattendue sur les résultats.

Ces travaux sur l'influence des observations dans la MRS nous ont conduit à proposer de tenir compte de cet aspect séquentiel dans l'inférence elle-même. L'utilisation de poids selon un schéma fixe semblant exclue devant la baisse de performance qu'ils impliquent en l'absence d'observations inattendues, nous nous sommes inspirés des travaux de Hu et Rosenberger sur la vraisemblance pondérée [83, 84] et avons adapté le concept de vraisemblance pondérée pertinente à la MRS. Ce concept vise à diminuer l'influence de certains sujets, le caractère paramétrique des poids permettant en outre de moduler la pondération en fonction des observations. Comme l'indique Hu [83], "Ces poids expriment la perception qu'a le statisticien des relations entre les observations et sont choisis habituellement sur des critères intuitifs". Ils sont donc le reflet de notre analyse de la MRS et explique le choix d'une fonction $\log\log$ abaissant potentiellement fortement le poids des premières observations en laissant les dernières avec des poids du même ordre. Nous avons montré l'intérêt d'une telle approche quelle que soit la probabilité cible et en présence ou non d'observations inattendues. Il est évident que d'autres fonctions de poids pourraient être envisagées ainsi qu'une extension à une inférence bayesienne.

L'observation d'une donnée inattendue au tout début d'un essai pose bien sûr également la question de la poursuite de cet essai et ce d'autant plus s'il s'agit d'un essai de phase I et que l'observation est une toxicité. En pratique, il est clair que l'essai sera interrompu d'autant

plus volontiers que la toxicité dose-limitante observée est sévère, voire qu'il s'agit d'un décès. Nous nous sommes donc limités volontairement au contexte des essais de phase II, pour lesquels l'observation d'un échec dès le premier patient est moins problématique (comme l'a démontré la conduite de l'essai réel). Néanmoins, dans le cadre d'un essai avec inclusions groupées, il est cependant moins clair que l'observation d'une toxicité conduise au même choix : le schéma *'3+3'*, encore largement utilisé, conduit d'ailleurs à re-administrer la même dose à 3 patients supplémentaires. Cependant, même dans cette situation, l'influence de cette première réponse inattendue est élevée. Nous préconisons donc l'utilisation de cette MRS modifiée en cas d'essai de phase I ou II avec inclusions groupées.

Chapitre 5

Conclusion

Au cours de cette thèse, nous avons montré l'intérêt des mesures d'influence pour la compréhension des modèles statistiques et plus particulièrement dans le cadre des modèles utilisés pour l'analyse de risques en compétitions et des modèles utilisés lors d'essais séquentiels de phase précoce. L'objectif n'était pas tant de développer des outils de mesure de l'influence individuelle des observations pour ces modèles (qui n'en possédaient pas), que de mieux appréhender leurs particularités en essayant de comprendre quelles observations étaient rendues déterminantes par l'utilisation même du modèle.

Nous avons ainsi proposé une mesure d'influence locale pour le modèle de Fine et Gray dérivée de celle proposée par Cook [3]. Elle nous a permis d'illustrer les différences qu'ils existent du point de vue de l'importance donnée aux dernières observations entre une modélisation du risque spécifique d'événement et une modélisation du risque associé à la fonction de sous-répartition de l'événement. Cette mesure se contente, néanmoins, de fournir la valeur de l_{max} qui ne représente que la "direction" de plus grande perturbation du modèle. Même si ce travail ne traitait pas des observations aberrantes (*outliers*), l'application la plus directe du développement proposé reste leur détection après la réalisation d'une analyse. Pour améliorer cette capacité de détection, une extension possible de ces travaux consiste à implémenter pour le modèle de Fine et Gray la mesure agrégée basée sur l'ensemble des vecteurs propres proposée par Poon et Poon [22]. De même, on pourrait envisager dans le cadre des modèles de survie de faire varier le seuil de détection des données aberrantes en fonction du rang, puisque nous avons montré que leur influence est intrinsèquement liée à leur rang.

Reste le problème de l'attitude à adopter face à ces observations. S'il semble raisonnable dans un premier temps de vérifier l'exactitude des données, les avis sont plus partagés quant au retrait éventuel de ces *outliers* [88]. En recherche clinique et plus particulièrement lors de la réalisation d'un essai thérapeutique, la situation est relativement claire : "en intention de traiter", on ne peut retirer les observations semblant aberrantes. Les méthodes statistiques doivent alors s'en accommoder et ne doivent pas être sérieusement perturbées par la présence de ces données.

Cette attitude se retrouve dans la deuxième partie de cette thèse. Partant d'un exemple réel, nous avons proposé une méthode pour mettre en évidence l'influence des premières observations. Loin des techniques développées pour les modèles de régression, nous nous sommes inspiré grandement des solutions mises en place pour des analyses de sensibilité en procédant par simulation. Ces simulations ont montré les conséquences dramatiques que peuvent avoir l'observation au début de l'essai d'une réponse inattendue sur la conduite d'un essai séquentiel de recherche de dose utilisant la méthode de réévaluation séquentielle. par ailleurs, indépendamment de l'observatino d'une réponse inattendue, et soupçonnant que l'aspect séquentiel de la MRS impliquait une influence décroissante avec le rang, nous avons proposé d'utiliser la méthode de vraisemblance pondérée pertinente. Cette méthode en diminuant l'influence des premières observations nous a permis de rendre plus robuste la MRS. Il reste cependant à l'adapter au cadre bayesien et ce d'autant plus qu'il existe une parenté entre le choix intuitif des poids [83] et le choix de la distribution *a priori* du paramètre dans l'inférence bayesienne.

Au total, cette "grille de lecture" que nous a fourni l'étude de l'influence individuelle, nous aura permis de mieux comprendre deux modèles fort différents. S'il est sans doute illusoire d'espérer que chaque analyse donne lieu à une étude d'influence individuelle, on ne peut que recommander son utilisation pour appréhender la complexité et les implications de certains modèles statistiques. Les influences de ces modèles parfois structurelles doivent être gardées à l'esprit, en particulier lors de l'interprétation des paramètres de ces modèles en épidémiologie clinique.

Bibliographie

[1] G.E.P. Box and N.R. Draper. *Empirical Model-Building and Response Surfaces.* Wiley Series in Probability and Mathematical Statistics. Applied probability and Statistics. John Wiely & Sons, Ltd., Chichester, 1987.

[2] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression.* Chapman and Hall, London, 1982.

[3] D. Cook. Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, 48 :133–169, 1986.

[4] G.A. Barnard. Discussion of professor Box's paper. *Journal of the Royal Statistical Society A*, 143 :404–406, 1980.

[5] L.A. Escobar and W.Q Meeker. Assessing influence in regression analysis with censored data. *Biometrics*, 48 :507–528, 1992.

[6] C.H. Chen and P.C. Wang. Diagnostic plots in Cox's regression model. *Biometrics*, 47 :841–850, 1991.

[7] D. Cook. Detection of influential observations in linear regression. *Technometrics*, 19 :1–12, 1977.

[8] D. Pregibon. Logistic regression diagnostics. *Annals of Statistics*, 9 :705–724, 1981.

[9] K.C. Cain and N.T. Lange. Approximate case influence for proportional hazards regression model with censored data. *Biometrics*, 40 :493–499, 1984.

[10] D. Collett. *Modelling survival data in medical research.* Chapman and Hall, second edition, 2003.

[11] D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34 :187–220, 1977.

[12] J.P. Fine and R.J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94 :496–509, 1999.

[13] J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method : a practical design for phase I clinical trials in cancer. *Biometrics*, 46 :33–48, 1990.

[14] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[15] F.E. Harrell. *Regression modeling strategies*. Springer, New York, 2001.

[16] D.A. Belsley, E. Kuh, and R.E. Welsch. *Regression diagnostics*. John Wiley, Hoboken, New Jersey, 1980.

[17] D. Pregibon. Resistant fits for some commonly used logistic models with medical pplications. *Biometrics*, 38 :485–498, 1982.

[18] D.A. Williams. Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, 36 :181–191, 1987.

[19] A.J. Lawrance. Deletion and masking in regression. *Journal of the Royal Statistical Society B*, 57 :181–189, 1995.

[20] W. thomas and R.D. Cook. Assessing influence on regression coefficients in generalized linear models. *Biometrika*, 76 :741–749, 1989.

[21] H.T. Zhu and Lee S.Y. Local influence for generalized linear models. *The Canandian journal of statistics*, 31 :293–309, 2003.

[22] W.Y. Poon and Y.S. Poon. Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical society. Series B (Statistical methodology)*, 61 :51–61, 1999.

[23] A.N. Pettitt and I. Bin Daud. Case-weighted measures of influence for proportional hazards regression. *Applied Statistics*, 38 :51–67, 1989.

[24] X. Wu and Z. Luo. Second-order approach to local influence. *Journal of the Royal Statistical Society B*, 55 :929–936, 1993.

[25] H. Zhu and H. Zhang. A diagnostic procedure based on local influence. *Biometrika*, 91 :579–589, 2004.

[26] E. Lesaffre and G. Verbeke. Local influence in linear mixed models. *Biometrics*, 54 :570–582, 1998.

[27] E. Demidenko and T.A. Stukel. Influence analysis for linear mixed-effects models. *Statistics in Medicine*, 24 :893–909, 2005.

[28] K. Van Steen, G. Molenberghs, G. Verbeke, and H Thijs. A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling : an international journal*, 1 :125–142, 2001.

[29] M Galea-Rojas, H. Bolfarine, and M. De Castro. Local influence in comparative models. *Biometrical journal*, 44 :59–81, 2002.

[30] C. Hill, C. Com-Nougué, A. Kramar, T. Moreau, J. O'Quigley, R. Senoussi, and C. Chastang. *Analyse statistique des données de survie*. Flammarion, 1996.

[31] R.J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Annals of statistics*, 16 :1141–1154, 1988.

[32] M.S. Pepe. Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association*, 86 :770–778, 1991.

[33] T.A. Gooley, W. Leisenring, J. Crowley, and B.E. Storer. Estimation of failure probabilities in the presence of competing risks : New representations of old estimators. *Statistics in Medicine*, 18 :695–706, 1999.

[34] M. Crowder. *Classical competing risks*. Chapman and Hall/CRC, 2001.

[35] E. Azoulay, C. Adrie, A. De Lassence, F. Pochard, D. Moreau, G. Thiery, C. Cheval, P. Moine, M. Garrouste-Orgeas, C. Alberti, Y. Cohen, and J.F. Timsit. Determinants of postintensive care unit mortality : A prospective multicenter study. *Critical Care Medecine*, 31 :428–32, 2003.

[36] J.R. Le Gall, J. Klar, S. Lemeshow, F. Saulnier, C. Alberti, A. Artigas, and D. Teres. The logistic organ dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU scoring group. *JAMA*, 276 :802–10, 1996.

[37] J.R. Le Gall, S. Lemeshow, G. Leleu, J. Klar, J. Huillard, M. Rue, D. Teres, and A. Artigas. Customized probability models for early severe sepsis in adult intensive care patients. Intensive care unit scoring group. *JAMA*, 273 :644–50, 1995.

[38] J.R. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter stu. *JAMA*, 270 :2957–63, 1993.

[39] S. Lemeshow and J.R. Le Gall. Modeling the severity of illness of ICU patients. A systems update. *JAMA*, 272 :1049–55, 1994.

[40] S. Lemeshow, D. Teres, J. Klar, J.S. Avrunin, S.H. Gehlbach, and J. Rapoport. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *JAMA*, 270 :2478–86, 1993.

[41] W.A. Knaus, D.P. Wagner, E.A. Draper, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, A. Damiano, and al. The APACHE III prognostic system. Risk prediction of hospital mortality forcritically ill hospitalized adults. *Chest*, 100 :1619–36, 1991.

[42] D. Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69 :239–241, 1982.

[43] P.M. Grambsch and T.M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81 :515–526, 1994.

[44] N. Reid and H. Crépeau. Influence functions for proportional hazards regression. *Biometrika*, 72 :1–9, 1985.

[45] B.E. Storer and J. Crowley. A diagnostic for Cox regression and general conditional likelihoods. *Journal of the American Statistical Association*, 80 :139–147, 1985.

[46] W.E. Barlow. Global mesure of local influence for proportional hazards regression models. *Biometrics*, 53 :1157–1162, 1997.

[47] C.B. Parker and E.R. Delong. A diagnostic for cox regression with discrete failure-time models. *biometrics*, 56 :996–1001, 2000.

[48] N. Sartori, K. Thomaseth, and Salvan A. Local influence analysis when interfacing toxicokinetic and proportional hazard models. *Statistics in Medicine*, 23 :2399–2412, 2004.

[49] W.H. Wei and J.S. Su. Model choice and influential cases for survival studies. *Biometrics*, 55 :1295–1299, 1999.

[50] W.H. Wei and Kosorok M.R. Masking unmasked in the proportional hazards model. *Biometrics*, 56 :991–995, 2000.

[51] A. Latouche, R. Porcher, and S. Chevret. Sample size formula for proportional hazards modeling of competing risks. In C. Serrat, M.L. Calle, G. Gomez, and O. Julia, editors, *Proceedings of the First Barcelona Workshop on Survival Analysis*, pages 91–92, 2002.

[52] A. Nardi and M. Schemper. New residuals for Cox regression and their application to outlier screening. *Biometrics*, 55 :523–529, 1999.

[53] D.M. Hawkins. *Encyclopedia of Biostatistics*, volume 4, chapter Outlier, pages 3222–3226. Armitage, P. and Colton, T., 1998.

[54] S. Chevret. *Statistical methods dor dose-finding experiments*. John Wiley and Sons, Chichester, 2006.

[55] C. Gatsonis and J. B. Greenhouse. Bayesian methods for phase I clinical trials. *Statistics in medicine*, 11 :1377–1389, 1992.

[56] Y. Lin and W. J. Shih. Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. *Biostatistics*, 2(2) :203–15, 2001.

[57] W.J. Shih and Lin Y. Traditional and modified algorithm-based designs for phase I cancer clinical trials. In S. (Eds.) Chevret, editor, *Statistical Methods for dose-finding experiments*, pages 61–90. John Wiley & Sons, Chichester, 2006.

[58] M. Gezmu and N. Flournoy. Group up-and-down designs for dose-finding. *Journal of statistical planning and inference*, 136 :1749–1764, 2006.

[59] J. O'Quigley and S. Chevret. Methods for dose finding studies in cancer clinical trials : a review and results of a monte carlo study. *Statistics in medicine*, 10 :1647–1664, 1991.

[60] J. Whitehead and H. Brunier. Bayesian decision procedures for dose determining experiments. *Statistics in medicine*, 14 :885–893 ; discussion 895–899, 1995.

[61] P. F. Thall and K. E. Russell. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*, 54 :251–264, 1998.

[62] J. O'Quigley and L. Z. Shen. Continual reassessment method : a likelihood approach. *Biometrics*, 52 :673–684, 1996.

[63] D. Faries. Practical modifications of the continual reassessment method for phase I cancer clinical trials. *J Biopharm Stat*, 4 :147–164, 1994.

[64] E. L. Korn, D. Midthune, T. T. Chen, L. V. Rubinstein, M. C. Christian, and R. M. Simon. A comparison of two phase I trial designs. *Statistics in medicine*, 13 :1799–1806, 1994.

[65] S. N. Goodman, M. L. Zahurak, and S. Piantadosi. Some practical improvements in the continual reassessment method for phase I studies. *Statistics in medicine*, 14 :1149–1161, 1995.

[66] D.M. Potter. Phase I studies of chemotherapeutic agents in cancer patients : A review of the designs. *Journal of Biopharmaceutical statistics*, 16 :579–604, 2006.

[67] S. Chevret. The continual reassessment method in cancer phase I clinical trials : a simulation study. *Statistics in medicine*, 12 :1093–1108, 1993.

[68] C. Ahn. An evaluation of phase I cancer clinical trial designs. *Statistics in medicine*, 17 :1537–49, 1998.

[69] S. Moller. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Statistics in medicine*, 14 :911–922 ; discussion 923, 1995.

[70] J.M. Heyd and B.P. Carlin. Adaptative design improvement in the continual reassessment method for phase I studies. *Statistics in medicine*, 18 :1307–1321, 1999.

[71] S. Piantadosi, J. D. Fisher, and S. Grossman. Practical implementation of a modified continual reassessment method for dose-finding trials. *Cancer Chemother Pharmacol*, 41 :429–36, 1998.

[72] J. O'Quigley and E. Reiner. A stopping rule for the continual reassessment method. *Biometrika*, 85 :741–748, 1998.

[73] S. Zohar and S. Chevret. The continual reassessment method : Comparison of bayesian stopping rules for dose-ranging studies. *Statistics in medicine*, 20 :2827–2843, 2001.

[74] J.M. Treluyer, S. Zohar, E. Rey, P. Hubert, F. Iserin, M Jugies, R. Lenclen, S. Chevret, and G Pons. Minimum effective dose of midazolam for sedation of mechanically ventilated neonates. *Journal of Clinical Pharmacy and Therapeutics*, 30 :479–485, 2005.

[75] M. de Spirlet, J.M. Treluyer, S. Chevret, E. Rey, M. Tournaire, D. Cabrol, and G. Pons. Tocolytic effects of intravenous nitroglycerin. *Fundamental and clinical pharmacology*, 18 :207–213, 2004.

[76] L. Desfrere, S. Zohar, P. Morville, A. Brunhes, S. Chevret, G. Pons, G. Moriette, E. Rey, and J.M. Treluyer. Dose-finding study of ibuprofen in patent ductus arteriosus using the continual reassessment method. *Journal of clinical pharmacy and therapeutics*, 30 :121–132, 2005.

[77] E. Fabre, S. Chevret, J. F. Piechaud, E. Rey, F. Vauzelle-Kervoedan, P. D'Athis, G. Olive, and G. Pons. An approach for dose finding of drugs in infants : sedation by midazolam studied using the continual reassessment method. *British journal of clinical pharmacology*, 46 :395–401, 1998.

[78] F. Lefrere, S. Zohar, J.L. Bresson, S. Chevret, A. Mogenet, F. Audat, I. Durand-Zaleski, D. Ghez, L. Dal Cortivo, P. Piesvaux, M. Cavazzana-Calvo, and B. Varet. A double-blind low dose-finding phase II study of granulocyte colony-stimulating factor combined with chemotherapy for stem cell mobilization in patients with non-hodgkin's lymphoma. *Haematologica*, 91 :550–553, 2006.

[79] J.B. Copas. Binary regression models for contaminated data. *Journal of the royal statistical society. Series B*, 50 :225–265, 1988.

[80] G. B. Wetherill. Bayesian decision procedures for dose determining experiments. *Journal of the Royal Statistical Society B*, 25 :1–48, 1963.

[81] G. B. Wetherill and Glazebrook K.D. *Sequential methods in statistics*. Chapman an Hall, London, third edition, 1986.

[82] V. Roshan Joseph. Efficient Robbins-Monro procedure for binary data. *Biometrika*, 91 :461–470, 2004.

[83] F. Hu. The asymptotic properties of the maximum relevance weighted likelihood estimators. *The Canadian journal of statistics*, 25 :45–49, 1997.

[84] F. Hu and J.V. Zidek. The weighted likelihood. *The Canadian journal of statistics*, 30 :347–371, 2002.

[85] F. Hu and W. F. Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials*, chapter 7, pages 107–119. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, New Jersey, 2006.

[86] Y. K. Cheung and R. Chappell. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*, 56 :1177–82, 2000.

[87] J. O'Quigley. Retrospective analysis of sequential dose-finding designs. *Biometrics*, 61 :749–756, 2005.

[88] V. Barnett and T. Lewis. *Outliers in statistical data*. John Wiley, Chichester, third edition, 1994.